Signature:

______________________________ _______________

Qing He Date

Machine Learning Methods in Large Scale Neuroimaging Study

By

Qing He
Doctor of Philosophy

Biostatistics

Jian Kang, Ph.D.
Advisor

Tianwei Yu, Ph.D.
Co-Advisor

Zhaohui Qin, Ph.D.
Committee Member

Jun Kong, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Machine Learning Methods in Large Scale Neuroimaging Study

by

Qing He
M.S., Virginia Polytechnic Institute and State University, 2010
B.S., Beijing Normal University, 2006

Co-Advisor: Jian Kang, Ph.D.
Co-Advisor: Tianwei Yu, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2017

ABSTRACT

Machine Learning Methods in Large Scale Neuroimaging Study

by

Qing He

The focus of this dissertation is on developing machine learning methods for analysis of the large-scale neuroimaging data. It consists of three topics.

In the first topic, we develop a spatial-temporal Gaussian process regression (STGPR) model for Bayesian analysis of longitudinal imaging data. Our goal is to study progressions of the brain activities in different brain regions and how they are associated with time-independent predictors (disease status, gender, etc.) and time-varying predictors (age, weight, etc.). We assign Gaussian processes priors to spatial-temporal varying coefficients in the model. To cope with the large-scale dataset, we develop three fast posterior computation algorithms based on the Karhunen-Loeve expansions on the Gaussian processes. Compared with a voxel-wise linear model approach, we demonstrate the advantages of the proposed method in a simulation study, where we propose two metrics: relative L1 loss and gradients relative L1 loss for measuring coefficient estimation accuracy. We apply the proposed method to the analysis of the longitudinal positron emission tomography (PET) data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study and obtain some meaningful results.

In the second topic, we use ensemble classification methods to predict disease status using neuroimaging data as biomakers in clinical studies. According to the existing brain atlas, the whole brain can be partitioned into many brain regions. For each brain region, we use voxel-level brain image to generate important classification

features, using which we develop many region-level basic classifiers. Then we combine those basic classifiers through linear programming boost (LPBoost) to find an optimal feature combination rule for classification. We develop an efficient column generation algorithm to solve both binary and multi-class LPBoost problem in high-dimensional feature space. We show the proposed method can improve the performance of basic support vector classifiers (SVC) dramatically and outperform other existing alternatives. We use the proposed method to analyze a large-scale resting state fMRI data in the Autism Brain Imaging Data Exchange (ABIDE) study data, leading to a better prediction accuracy than the existing best result.

In the third topic, we make Bayesian inference on peaks of smooth curves in a nonparametric regression model, where we determine the peak location based on gradients of the curve. We assign a Gaussian process prior to the smooth curve of interest. We show that the joint posterior distribution of the curve, its first derivative and the second derivative follow a multivariate Gaussian process. This result leads a straightforward posterior inference on peak locations and magnitudes. In the simulation study, we demonstrate that the proposed peak identifier outperforms the existing non-parametric kernel smoothing method in different scenarios. We apply the proposed method to analysis of electroencephalogram (EEG) time series in a study of alcoholism. In particular, the proposed method is applied to find the peaks of the EEG time series in the temporal domain and peaks of the signal power in the frequency domain. We construct a peak-based classifier on alcoholism versus normals, which achieves a 80% classification accuracy.

Machine Learning Methods in Large Scale Neuroimaging Study

by

Qing He
M.S., Virginia Polytechnic Institute and State University, 2010
B.S., Beijing Normal University, 2006

Co-Advisor: Jian Kang, Ph.D.
Co-Advisor: Tianwei Yu, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2017

ACKNOWLEDGEMENTS

I am grateful for a number of people who have been invaluable for my completion of this dissertation.

First and foremost, I want to thank Dr. Jian Kang, for his continuous support and guidance for my Ph.D. study and research. During the past seven years, he has constantly provided me timely advice, endless patience, and insightful suggestions to overcome different challenges in my research. The breath and depth of his knowledge, his inspiring ideas, and the algorithm developing skills he has taught me made my Ph.D. experience productive and stimulating. It has been a great honor for me to have him as my adviser.

I would like to express my heartfelt gratitude to Dr. Tianwei Yu and Dr. Zhaohui Qin who guided me on several research projects that brought me into machine learning and Bayesian computation research area. They set up role models and showed me the type of researchers I would like to be. As my dissertation co-adviser and committee member, they provided very valuable inputs in my dissertation research. I am also extremely grateful to my committee member Dr. Jun Kong for his constructive suggestions, which significantly improved the quality of my work.

Many thanks to the department faculties especially Dr. Qi Long, Dr. Ying Guo, Dr. Amita Manatunga, and Dr. Lance Waller. Not only I learned a lot from their areas of expertise in Biostatistics, but I also benefited from their kind consultation in research and career path. Many thanks to staff members especially Melissa and Mary for making the department a family like environment. Many thanks to my supportive

schoolmates especially Yunxuan Jiang and Jie Chen for being my best friends and sharing my important moment both in school and in life.

Last, I would like to thank my parents for their unconditional love and support, my husband for his constant belief in me, and my son Orville for being my sunshine and giving me strength. I feel extremely lucky to have my family by my side.

TABLE OF CONTENTS

ABSTRACT . . . .

ACKNOWLEDGEMENTS ii

LIST OF FIGURES vii

LIST OF TABLES ix

CHAPTER

I. Introduction 1

1.1 Overview 1
1.2 Human Brains and Functional Neuroimaging Data 2
1.2.1 Basic Knowledge of Human Brain 2
1.2.2 Positron Emission Tomography (PET) Imaging . . . 5
1.2.3 Functional Magnetic Resonance Imaging (fMRI) . . 6
1.2.4 Electroencephalography (EEG) of Brain Electrical Activity 7
1.3 Statistical Methods of Neuroimaging Studies 8
1.3.1 Preprocessing Pipeline 9
1.3.2 Activation Studies 10
1.3.3 Connectivity Analysis 11
1.3.4 Prediction and Classification 13
1.4 Motivation Example and Proposed Research 14
1.4.1 Alzheimer's Disease Neuroimaging Initiative Study (ADNI) 14
1.4.2 Autism Spectrum Disorder (ASD) Study 15
1.4.3 Electroencephalogram Alcoholism Study 16

II. Topic 1: Bayesian Gaussian Process Modeling of Large Scale Longitudinal Neuroimaging Data 17

2.1 Introduction 17

2.1.1 Longitudinal Neuroimaging Analysis 17
2.1.2 Machine Learning methods for Neuroimaging Study 19
2.1.3 Gaussian Processes and Its Properties 21
2.2 The Model 23
2.2.1 A Spatial-Temporal Model 23
2.2.2 Prior Specifications 23
2.2.3 Model Representation 25
2.2.4 Posterior Computation 27
2.2.5 An STGPR based Classifier 29
2.3 Simulation Studies 30
2.4 ADNI Data Analysis 34
2.5 Discussion 37
2.6 Appendix 38
2.6.1 Explicit Forms of Eigen Values and Eigen Vectors . 38
2.6.2 Full Conditional Posterior Density of Hyperparameters 40

III. Topic 2: Ensemble Classification Methods For Feature Combination of Large Scale Neuroimaging Data 42

3.1 Introduction 42
3.1.1 Disease Prediction Using Neuroimaging Data . . . 44
3.1.2 Support Vector Machine Classifiers 47
3.1.3 Ensemble Classification Methods 48
3.2 Methods 50
3.2.1 Binary Support Vector Classifiers 52
3.2.2 Binary LPBoosting through Column Generation . . 56
3.2.3 Multiclass SVM and LPBoost 59
3.3 Simulation Study 63
3.4 ABIDE Data Analysis 71

IV. Topic 3: Bayesian Nonparametric Inference on Peaks via Spatially Adaptive Non-stationary Gaussian Processes 80

4.1 Introduction 80
4.2 Methods 82
4.2.1 Posterior Computation 84
4.3 Simulation Study 86
4.3.1 Effect of Noise 86
4.3.2 Effect of Multiple Equilibrium Points 87
4.4 EEG Data Analysis 87
4.4.1 Peak of Alpha Band in Temporal Domain 90
4.4.2 Peak of Power in Frequency Domain 91
4.4.3 Classification based on Peak and Magnitude 92
4.5 Appendix 97

BIBLIOGRAPHY 102

LIST OF FIGURES

Figure

2.1 Maps of coefficients' effects estimated by proposed STGPR for ADNI data. Plots include population baseline effects, $\hat{\alpha}_t$, spatial effects for time-varying covariates, $\hat{\beta}_j$, and spatial-temporal effects for time-independent covariates, $\hat{\eta}_{t,j}$ 36

3.1 One replicate of simulated imaging data of three disease groups, containing 100 regions and each has a 10×10 voxel space, the simulated images are generated under senario I, so that the data has no spatial correlation 65

3.2 One replicate of simulated imaging data of three disease groups, containing 100 regions and each has a 10×10 voxel space, the simulation images are generated under senario II 66

3.3 One replicate of simulated imaging data of three disease groups, containing 100 regions and each has a 10×10 voxel space, the simulation images are generated under senario III 67

4.1 Plots of original EEG signal and the filtered alpha wave on electrode P6 of subject "co2c0000337" (first row) from control group and subject "co2a0000364" from alcohol group. Data collected from different stimulus types are shown in three columns 91

4.2 Plots of the fitted curve (blue line) for the filtered alpha wave and identified peak location (red triangle) by the proposed method. The example data is the alpha wave from electrode CP2 of subject "co2c0000337" (first row) from control group and subject "co2a0000364" from alcohol group and across different stimulus types 92

4.3 Plots of the fitted curve (blue line) for the power curve of alpha wave and identified peak location (red triangle) by the proposed method. The example data is the power of frequency of the alpha wave on electrode P6 of subject "co2c0000337" (first row) from control group and subject "co2a0000364" from alcohol group and across different stimulus types 95

4.4 Original EEG signal curves for all control subjects in the first row and alcoholic subjects in the second row at electrode P6, averaged for all trials under each of the three experiments 97

4.5 Filtered Alpha wave signals for all control subjects in the first row and alcoholic subjects in the second row at electrode P6, averaged across all trials under each of the three experiment 98
4.6 Alpha wave power curves for all control subjects in the first row and alcoholic subjects in the second row at electrode P6, averaged across all trials for each of the three experiments 99

LIST OF TABLES

Table

2.1 The hyperprior parameters used in simulation study to generate baseline function, coefficients, and random noise 31

2.2 Simulation results comparing LM and STGPR in regards to estimation and prediction of both coefficients and outcome variables through RL_1 Loss(relative L_1 Loss) and GRL_1 (gradients' relative L_1 Loss) averaging over space and necessarily over time. For STGPR, Algorithm II is used on simulated data of size 60×60 and Algorithm III is used for simulated data of size 100×100. The values in the parenthesis are standard error of the quantities of interest. 33

2.3 A summary of results for ADNI data analysis. MSE(mean squared error) and PMSE(predicted mean squared error) are provided based on five-fold cross validation conducting on both Hippocampus left and right regions and for all subjects, AD group, and MCI group respectively. 35

3.1 Simulation Setting of Mean and Standard Deviation Values of 100 Regions 64

3.2 Three Scenarios of Simulation Study 64

3.3 Result summary of simulation study for binary classification. The test errors of basic classifier SVM, PCA+SVM, and LPBoost methods are listed based on one random Train/Valid/Test split 68

3.4 Result summary of simulation study for three group classification. The test errors of basic classifier SVM, PCA+SVM, and LPBoost methods are listed based on one random Train/Valid/Test split . . . 68

3.5 Result summary of simulation study for binary classification. The relative importance of different regions (one random Train/Valid/Test split) based on basic SVM classifiers and on LPBoost are listed. . . 69

3.6 Continued to 3.5 70

3.7 Simulation results for binary classification and three group classification comparing LPBoost method with PCA+SVM method and regional SVM methods based on 50 repeated experiments, randomly spliting train, valid, test data sets. The measurement metrics includ classification error and its standard error, classification error ratio between two methods of interest 71
3.8 A summary of stratified random sampling conducted for ABIDE data in experiment I 73
3.9 A summary of Train/Test/Valid split of ABIDE data in experiment II 74
3.10 A summary of Train/Test/Valid split of ABIDE data in experiment III 75
3.11 LPBoost classification error under three experiments. Results for experiment I is based on 50 replicates of stratified random sample of Train/Valid/Test data sets 75
3.12 A result summary of relative importance of different brain regions for ABIDE analysis based on 50 repeated stratified random samples under Experiment I 76
3.13 Continued to 3.12 77
3.14 A result summary of relative importance of different brain regions for ABIDE analysis based on Experiment II 78
3.15 A result summary of relative importance of different brain regions for ABIDE analysis based on Experiment III 79
4.1 Simulation results for the proposed method compared to the nonparametric kernel smoothing (NKS) method in regards to mean squared error (MSE) of peak location and 95% confidence interval coverage . 86
4.2 Simulation results for proposed method compared to NKS method in regards to MSE of peak location and 95% confidence interval coverage for two peak locations respectively 87
4.3 A list of electrodes that show significant difference in time locations of the first peak after stimulus on alpha wave. The hypothesis test is based on the Student's T test and exams the quantity of interest between control group and alcohol group under three different experiments 93
4.4 A list of electrodes that show significant difference in time locations of the highest peak after stimulus on alpha wave. The hypothesis test is based on the Student's T test and exams the quantity of interest between control group and alcohol group under three different experiments 93
4.5 A list of electrodes that show significant difference in average interval of peak locations on alpha wave. The hypothesis test is based on the Student's T test and exams the quantity of interest between control group and alcohol group under three different experiments 94
4.6 A list of electrodes that show significant difference in peak alpha frequency. The hypothesis test is based on the Student's T test and exams the quantity of interest between control group and alcohol group under three different experiments 95

4.7 EEG classification results based on three experiments presented in different rows and four types of location related predictors presented in different columns. The mean classification error and its corresponding standard error (in parenthesis) are calculated based on tenfold cross validation 96
4.8 A list of EEG data for alcoholism study in regards to the total number of trials of each subjects under three experiments 100
4.9 Continuous from Table (4.8) 101

CHAPTER I

Introduction

1.1 Overview

The rising of neuroimaging technology has provided various tools to study brain activities. It improves our understanding of both the neurophysiology of healthy individuals and the pathophysiology of patients suffering from mental illnesses or major psychiatric disorders, e.g. Alzheimer's disease, Autism, Alcoholism. Common methods of functional neuroimaging, e.g. positron emission tomography (PET), functional magnetic resonance imaging (fMRI), electroencephalography (EEG), provide quantitative measurements of brain activities, either through three dimensional (3D) images on hundreds of thousands of cuboid elements, called "voxels", or through external sensors attached to specific locations of the head. All types of brain imaging data are subject to common properties such as high dimensionality, complex spatial correlation structure, and intrinsic temporal associations, which pose challenges for statistical modeling. Therefore, our main objective is to develop new statistical methods to make inference on neuroimaging data and apply them for clinical purpose of disease diagnosis and treatment selection.

This dissertation is organized as follows: the remainder of Chapter I provides background information on the topics in neuroimaging researches, current statistical methods of neuroimaging analyses, motivating examples, as well as outlines our proposed research. Chapter II presents a spatial-temporal Gaussian process regression model for longitudinal neuroimaging data and its application to Alzheimer's Disease Neuroimaging Initiative Study (ADNI). Chapter III presents an ensemble classifier using linear programming boost for feature combination of large scale neuroimaging data and applies to a study of Autism Brain Imaging Data Exchange (ABIDE). Chapter IV presents a new Bayesian nonparametric method for making inference on peaks of curves. We demonstrate this method in the analysis of EEG data in an alcoholism study.

1.2 Human Brains and Functional Neuroimaging Data

1.2.1 Basic Knowledge of Human Brain

The human brain, as the most complex organ, produces feeling, emotion, and memory, and coordinates body actions while perceiving the outside world. The average weight of adult human brain is about 1.4 kg, containing about 86 non-neural cells (glial) and 85 billion neuron cells. There are two types of brain tissue defined anatomically including grey matter (cerebrum cortex) which consists of neuronal cell bodies, neuropil, glial cells and capillaries, and white matter which mostly contains myelinated axons as tracts to interconnect different regions of the cerebral cortex and supporting structure.

Human brain has its parcellation and atlases. The human brain contains the brainstem, cerebellum, and cerebrum (neocortex). The cerebrum responds to higher order reasoning, learning, and personality and is our major research interest. The cerebrum consists of two hemispheres (right and left) connected by white matter commissural fibers (e.g. corpus callosum). Each cerebral hemisphere is conventionally divided into four lobes: frontal, parietal, temporal, and occipital. The two hemispheres and four lobes provide us a general map of human brain. For population brain imaging studies, the individual brain images are usually first normalized into a common coordinate space to accommodate the between subject variation of brain size and orientation. The Talairach space and the Montreal Neurological Institute (MNI) space are the two most widely used atlas spaces. The Talairach coordinate system is based on a 4 stereotaxic atlas of the human cerebral cortex published by *Tzourio-Mazoyer et al.* (2002). Each brain location is defined by three dimensional coordinates with its distance from the midpoint of a brain white matter structure called the anterior commissure. The atlas is built based on a single brain of a 60-year-old French woman with mental disorder. Despite of its popularity, the Talairach space is quite different from normal brains. For example, the Talairach brain is considerably smaller than the average brain by up to 10 millimeters in each dimension. Later, the MNI coordinate system defines a new standard brain by using a large series of MRI scans of healthy normal controls (*Evans et al.*, 1993). These atlases differ in shape and size, and have been installed in common neuroimaging processing software. After normalization, finer cerebral cortex parcellation can be constructed and is desired for in-depth study. One fine cerebral cortex parcellation is defined by Brodman areas (48 regions), which are based on cytoarchitecture, or organization of cells. The Automated Anatomical Labeling (AAL) regions (116 regions) are constructed through the identification of major and minor sulci/gyri on a T1 MRI with subsequent labeling based on anatomical location (*Tzourio-Mazoyer et al.*, 2002), which is more used for functional neuroimaging-based research.

Brain connectivity and signal transmitting represent the status and function of a human brain. There are up to tens of thousands of neurons that passing signals between each other via synapses. The pattern and strength of such connections keep changing for every second of our lives according to updated experience, learning, and reinforcement. The changes determine the brain functions such as memory, personality, and habit. Therefore, the brain structure is not only shaped by genes but also even more by experience. There are various ways to pass signals between neurons, e.g. by electronic, magnetic and chemical pathways. In synapse, signals are passed between neurons by releasing and capturing neurotransmitters such as dopamine, acetylcholine, and serotonin. The neurotransmitters are very important for brain activity, and abnormality of them is related to diseases. For example, a deficiency in serotonin in limbic system is linked to depression or mood disorders. When building such electronic or chemical signal passing channels, energy is needed, which is provided by glucose and oxygenated-haemoglobin (generating APT). Therefore, high level brain region activity is usually synchronized with higher metabolite rate and glucose and oxygenated-haemoglobin concentration. All of the above are considered as the fundamental signals that neuroimaging research relies on to study human brains.

Neuroimaging is traditionally divided into structural and functional imaging. Structural imaging maps the brain anatomy and includes computed tomography (CT) and MRI. Functional imaging seeks to examine the physiological properties of the brain, either at rest or during task-induced activation. There are a variety of methods that maps human brain functioning. For example, PET and fMRI measure localized changes in cerebral blood flow, which is also referred to as activations. The two neuroimaging maps are able to show neuronal activity with relatively high spatial resolution ($\leq$ 1mm), but the temporal resolution (2-20 sec) is limited by the much slower rate of brain regional blood flow and blood oxygenation. In contrast, techniques such as electroencephalography (EEG) and magnetoencephalography (MEG) map the underlying electrical activity of the brain cortex. These methods allow high temporal resolution of neural processes, but have poor spatial resolution (over 1 cm). While each modality is interesting in its own right, in this dissertation we focus and introduce new statistical methods applied to fMRI, PET data, and EEG. In the following three sections, we give a brief description of the main principles on which these three neuroimaging techniques are based. In this dissertation we concentrate on how they are applied to human brain neuroimaging.

1.2.2 Positron Emission Tomography (PET) Imaging

PET is a nuclear imaging technique for mapping brain function or other molecular processes in the body. It requires injection or inhale of radioactively labeled chemicals into a subject's bloodstreams. The labeled compound, also called radiotracer, goes to the areas of the brain and body that use it while the subject is engaged in some type of mental or physical activity. Sensors in the PET scanner then measures emission related to the positron decay and generate images of the distribution of the radio chemicals throughout the brain and body. Different colors or degrees of brightness on a PET image represent different levels of tissue or organ function. This method hence provide a functional view of the brain and body. PET is not an exact measure of brain function since it depends upon certain assumption about what happens when an area of the brain becomes active. The assumptions include: 1) cerebral metabolism requires glucose metabolism, which requires oxygen from blood flow, i.e. there will be more blood flow in parts of the brain that are more active; 2) cerebral blood flow varies locally with corresponding local variations in neural activity.

PET has very high biochemical sensitivity and selectivity which allow probing the neurochemical processes at the molecular level. Its temporal and spatial resolution are inferior to that of fMRI. The mean free path of the positron in brain tissue limits the spatial resolution of PET scanning to about 4 mm. However, PET images can be superimposed on subject's MRI images, providing detailed information about specific brain areas involved in a wide variety of functions. Spatial resolution of PET data depends upon several other factors: the size and type of the crystal used in the scanner scintillator to detect the gamma radiation emission, the energy of the positron emitted etc. Temporal resolution depends mainly upon half-life of the isotope. Safety regulations require to wait 5 half-lives between injection the radioactive tracers. In a typical PET experiment, a brain scan is taken during a control task (e.g., resting with eyes closed). The control task is compared to brain scan taken while the subject is exposed to the experimental treatment or performing the experimental task. To determine the brain activity that can be attributed to the experimental condition, the difference between the PET scans is calculated.

1.2.3 Functional Magnetic Resonance Imaging (fMRI)

fMRI, as opposed to PET, is a relatively safe and non-invasive technique for generating maps of and studying brain activity. fMRI data consist of a 3D sequence of individual magnetic resonance images (MRI) which record a subject's brain activity. In general, MRI maps the water distribution in the brain which is based on an interaction between radio waves and atomic nuclei called nuclear magnetic resonance (NMR) (*Higgins*, 1996). To obtain an MRI, a subject is placed in a field of a large electromagnet (generally from 1.5 to 4.0 Tesla) that aligns the magnetization of hydrogen atoms in the brain. A hydrogen nucleus whose spin is oriented parallel to the applied magnetic field is said to be relaxed or in the low energy state, while a nucleus whose spin is oriented against the magnetic field is said to be in an excited or high energy state. When the scanner injects a pulse of radio frequency (RF), the nuclei is excited and raised out of their low energy states. When the RF pulse is removed, the hydrogen nuclei return to their original aligned position, and emit RF energy that measured by the scanner. The absorbing and emitting happens only when the frequency of the input radio waves equals the natural resonance frequency, i.e. the Larmor frequency, of the element in the magnetic field. Because of this, it is possible to highlight different characteristics of the imaged tissue by adding RF or gradient pulses and carefully choosing their timing. The useful contrast in MRI comes not only from spatial variation in the density of water, but also from differences in nuclear magnetic properties known as relaxation. They are characterized by distinct rates or relaxation times, used in MRI to distinguish between tissue types. Three relaxation times are of primary interest in MRI: T_1, T_2 and T_2^*. T_1 effects measure recovery of longitudinal magnetization that is parallel to the main magnetic field. T_2 refers to decay of transverse magnetization that is perpendicular to the main magnetic field. T_1 time refers to interval where 63% of longitudinal magnetization is recovered, and T_2 time refers to the interval where only 37% of original transverse magnetization is present. When T_2 dephasing is due to one or more localisable sources, it is referred to as as T_2^*. fMRI is a T_2^* image. The raw data obtained from an MRI scanner are collected in the frequency domain. The inverse Fourier transform is then used to transfer the data into image space, where data analysis is performed (*Ogawa et al.*, 1990).

1.2.4 Electroencephalography (EEG) of Brain Electrical Activity

Electroencephalography (EEG) is an electrophysiological monitoring method to record electrical activity of the brain. It requires to place multiple electrodes along the scalp and measure voltage fluctuations resulting from ionic current within the neurons of the brain. In clinical contexts, EEG is usually referred as the recording of the brain's spontaneous electrical activity over a period of time. Research based on EEG generally focus either on event-related potentials or on the spectral content of EEG. The former investigates potential fluctuations time locked to an event like stimulus onset or button press. The latter analyses the type of neural oscillations that can be observed in EEG signals in the frequency domain.

For clinical purpose, EEG is most often used to diagnose epilepsy, which causes abnormalities in EEG readings. Some other symptoms, i.e. diagnose sleep disorders, depth of anesthesia, coma, encephalopathies, and brain death can also be reflected in EEG data. EEG has used to be a first-line method of diagnosis for tumors, stroke and other focal brain disorders. Recently with the advent of high-resolution anatomical imaging techniques such as MRI and CT, the use of EEG has decreased. Despite limited spatial resolution, EEG continues to be a valuable tool for research and diagnosis. It is one of the few mobile techniques available and offers millisecond-range temporal resolution which is not possible with CT, PET or MRI.

1.3 Statistical Methods of Neuroimaging Studies

Exploring scientific and clinical outcomes have always been primary objectives for neuroimaging studies. The main focus of neuroimaging studies include the followings areas. 1) Activation studies target to identify particular brain regions that associate significant signal changes. The general objective of activation study is to localize brain areas that involved in task-related signal processing. Another objective is to compare response and non-response neural regions among different subgroups of individuals, i.e. treatment groups. 2) Connectivity studies seek to identify brain areas that show similar patterns of activity over time, yielding distributed networks of brain function. 3) Predictive studies try to use neuroimaging scans and patients' demographics to predict future neural outcomes. Prediction in neuroactivity presents influencing clinical significance, for example, studying the change in brain activities over time can help to monitor disease progression; forecasting brain reaction to different treatment strategies can provide insights on medical effect thus assist therapeutic designs; etc. 4) Classification study to classify individuals into different groups based on their neuroimaging biomarkers. Instead of considering the neuroimages as dependent factors to explore brain physiology, the imaging data can be utilized as features to predict clinical outcomes such as disease status and treatment response. The classification methods can also select neuroimaigng based biomarker for diagnostic purposes.

Data from a functional neuroimaging study consist of a series of 3D images, typically obtained while the subject performs a certain cognitive, behavioral or emotional task, or while at rest. In an fMRI study, typically hundreds of such 3D images are obtained, taken approximately 2-4 seconds apart. In a PET study, the number of scans is significantly smaller since the maximum number of scans per subject is limited due to the total isotope dose allowed. The brain images are obtained much less frequently due to the same reason. Each 3D image comprises of a large number (>100,000 in an fMRI study; usually smaller for PET) of voxels. In addition, the experiment may be repeated for the same subject, as well as for multiple subjects. Because of the neurophysiology of the network organization of the brain, spatial correlations are very likely. Also, temporal correlations both within and between scanning sessions are present due to repeated scanning.

1.3.1 Preprocessing Pipeline

The scans images usually undergo several preprocessing steps before statistical analysis in order to reduce artifacts caused by the scanner machine and possible movement of the subject, and to map individual brains to a standard brain atlas in order to perform group analysis and to make population inference.

For PET scan data, the pre-processing steps include the following steps. 1) Frame extraction: separate frames (usually five minutes apart) are extracted from the image file for registration purposes. 2) Co-registration: separate frames are co-registered to one another lessening the effects of patient motion. 3) Averaging: all five-minute apart frames are averaged into one image. 4) Spatial re-orientation and intensity normalization: each subjects image from their baseline PET scan is reoriented into a standard 16016096 voxel image grid having 1.5 mm cubic voxels. 5) Smoothing: each image set is filtered with a scanner-specific filter function to produce images of a uniform isotropic resolution. Details of the PET scan preprocessing procedure can refer to `http://adni.loni.usc.edu/methods/pet-analysis/pre-processing/`.

For fMRI data, the following steps are involved. 1) Brain extraction: strip the skull and non-brain tissues and generates a brain mask. 2) Motion correction: realigns 3D images to a common reference by using rigid body transformation with 6 degrees of freedom. 3) Slice timing correction: each voxel's time course is shifted by interpolation and resampling in order to be assumed that they are measured simultaneously. 4) Registration: register each subjects brain to a standard template brain atlas for example MNI space or Talairach space, using linear/affine transformation with 12 degree of freedom or nonlinear matrix transformation. 5) Normalization which is important in group analysis and attempts to register each subject's brain to a standard template brain; 6) Spatial smoothing: convolving the 3D images with a Gaussian kernel. Spatial smoothing is done for several reasons: it may improve inter-subject registration, it ensures that the assumptions of the random field theory are satisfied, and to increase signal to noise ratio. However, spatial smoothing causes a loss of acquired data and may introduce artificial spatial correlation between nearby voxels. The above preprocessing steps and the order in which they are performed are important because they influence both the spatial and temporal correlation structure of the image data.

For EEG data, detailed preprocessing steps are explained in Chapter (IV).

1.3.2 Activation Studies

The general objective of activation studies is to identify brain locations that are involved in the neural processing associated with tasks that subjects perform while in the scanner and possibly to compare these neural processing traits between tasks or between subgroups of individuals. Commonly, a two-stage statistical modeling procedure is used. At the first level, activation coefficients are calculated for single subjects under different experiment condition independently; then the second level conducts group level analysis on the summarized coefficients while adjusting for the covariates such as age, gender, and race (*Bowman et al.*, 2008). The general linear model (GLM) is the most population way to analyze the single subject fMRI data for activation study (*Worsley and Friston*, 1995; *Friston et al.*, 2002). Statistical hypothesis can be conducted to the model to test the increase in brain activity during the on signal versus during the off signal. Usually, individual model parameters can be tested using a t-test and subset of parameters using a partial F-test. Typically, multi-subject analysis of fMRI data from an activation study is performed using hierarchical models, which provide a framework for performing mixed-effects analysis. The group level or population parameters represents the effects related to different sessions or the effects associated with different subpopulations, e.g. treatment groups. Based on the hierarchical model, statistical inference between the different sessions or different groups can be conducted.

1.3.3 Connectivity Analysis

Functional connectivity has been defined as the temporal correlation between spatially remote neurophysiologic events. Since most brain functions are realized by certain kinds of functional connectivity of brain neurons, it is a very important focus of neuroimaging study. There are two research focus: 1) to study brain functional connectivity in resting state; 2) to study brain functional connectivity given specific stimuli. Usually during deep sleep the functional connectivity is much weaker than during awake status. Functional connectivity in resting state is much weaker than under stimuli. For fMRI data, functional connectivity analysis is based on a time series of BOLD signals. On the other hand, structural connectivity measures the white matter fibre tracts based on DTI imaging. The estimation is based on the path of water molecules' motion direction. It has intrinsic relation with functional connectivity because the neural signal is passed through axons in the white matter (*Honey et al.*, 2009). In the following, we mainly focus on the models for functional connectivity.

The idea of seed voxel approach for connectivity study start from selection of a voxel or a set of voxels based on functional or anatomical knowledge gained previously. Then the average time course of the selected voxel is studied and correlates with the remaining voxels. As a result the whole brain connectivities can be built based on the those reference voxels. The seed voxel approach is easy to calculate but the choice of the seed voxel could be subjective and it ignores network relationship between all voxels. *Greicius et al.* (2003) applied seed voxel approach to resting-state fMRI data for functional connectivity analysis, and revealed the connectivity between the posterior cingulate cortex and the ventral anterior cingulate cortex and provided evidence of a default mode networks (DMN) of brain function.

Cluster analysis is also widely used for connectivity analysis. It intends to identify networks or clusters that consist voxels of correlated patterns of measured brain activity. Although the clustering solutions only reflect functional association between voxels and do not define the underlying neuroanatomical connections, voxels from the same anatomical region ideally should exhibit high functional or spatial auto-correlation, validating that the neural responses within the same anatomical region are functional related. Most clustering methods are based on the dissimilarity of the activity time courses between different voxels. Two typical dissimilarity measures are Euclidean distance, which is for continuous variables, and Mahalonobis distances, which is for categorical variables. The voxels within the same cluster are expected to have more coherent performance. Commonly used clustering algorithms include hierarchical clustering and partitioning algorithms. There are also model-based clustering methods that do not require pre-specification of the number of clusters. For example, infinite mixture model via the Dirichlet Process or Chinese Restaurant Process allows simultaneous assignment of cluster membership along with optimization of the number of clusters.

There are also methods to use graph theory to study connectivity. The brain's structural and functional connectivity network has complex network topological features, such as high clustering, small worldness, the presence of high degree nodes, assortativity, modularity or hierarchy at both the whole-brain scale and local level. The graph composed by the nodes (voxels) and edges (connectivities between voxels) of brain network could provide an amount of metrics to describe the whole brain complex network, e.g. the degree of a node, assortativity, cluster coefficient, path length, connection density, and so on.

1.3.4 Prediction and Classification

Both activation and connectivity studies uses neuroimaging outcome to explore brain physiology. We can also consider the neuroimaging data as features to predict clinical outcomes such as disease status and treatment response (*Evans et al.*, 2006). To achieve this goal, feature selection and classifier construction are two major tasks.

Feature selection is a technique to choose a subset of most related features to the outcomes based on supervised learning models. Due to the curse of high-dimensionality, dimension reduction techniques are usually used for predictive analysis of neuroimaging data. There are two typical categories of feature selection algorithms. The first category is based on creating filters that preset threshold on statistic metrics of all features and elimitates those that do not pass. Many methods have been developed to control the false positive discovery rate for the large scale tests, for example local false discovery rate method by *Efron* (2005). The second category is called wrapper methods, which is objective function oriented and searches for the optimum set of possible features to achieve highest objective function value. One example is the recursive feature elimination algorithm by *Guyon et al.* (2002). The shrinkage method such as Lasso and elastic network could also be categorized as wrappers, since the variables selected are based on the penalized objective function *Efron et al.* (2004); *Zou and Hastie* (2005). The selected features could be used as inputs of the following classification models.

Machine learning classification methods have been successfully applied to neuroimaging data, for example, support vector machines (SVM), aritificial neural network classifiers, decision tree based algorithm such as CART and random forest, Bayesian classifiers, and the most recent convolutional neural network (CNN). Those supervised learning procedures usually include two steps: model training and prediction. In the training step, a part of the subjects are used to build the model with the objective of high classification rate with certain constraints. In the prediction step, the rest of the subjects are used to test based on the trained model for evaluation and future use. The cross validation procedures such as k-fold and leave-one-out could be applied.

1.4 Motivation Example and Proposed Research

1.4.1 Alzheimer's Disease Neuroimaging Initiative Study (ADNI)

Alzheimer's Disease Neuroimaging Initiative (ADNI) `http://www.loni.ucla.edu/ADNI/` is a large national project with a goal to develop biomarkers of Alzheimer's Disease (AD) in elderly subjects, to define the rate of progress of mild cognitive impairment (MCI) in Alzheimer's disease, and to provide a large database which will improve design of treatment trials.

In this dissertation, PET data of 69 typical controls subject (TC), 49 AD patients, and 117 MCI patients are obtained from the ADNI database and used to study the disease status and progression. The neuroimaging scans and all clinical covariates are collected longitudinally at baseline, six months and twelve months. The goals include: 1) to predict subject's follow-up (18 month) brain image outcome based on the existing brain images; 2) to classify or predict patients' disease status given existing brain image measurements; 3) to investigate different effects of clinical variables on brain activities and disease status.

In the first dissertation topic, a novel Bayesian non-parametric model is proposed and used to study Alzheimer's disease. The model is able to learn both spatial association and temporal changes of PET scans through multiple time sessions. The model is used to predict the follow-up imaging outcome, to identify changes in brain areas, and to investigate the spatial and temporal influence of different clinical factors. The method outputs smoothed brain signals and adjust for patients' individual factors and characteristics. Both time-varying and time-independent covariates are considered so that we can achieve both group specific influence, i.e. treatment groups, and task specific effect, i.e. task indicator.

1.4.2 Autism Spectrum Disorder (ASD) Study

Autism spectrum disorder (ASD) is a widely recognized disease characterized by qualitative impairment in social reciprocity, and by repetitive, restricted, and stereotyped behaviors. Due to its high prevalence in children with a more than 1% occurrence rate, there is strong need to further understand the mechanisms underlying ASD in order to identify ways of earlier diagnosis, optional treatment selection, and better outcome prediction. Autism Brain Imaging Data Exchange (ABIDE) is a consortium of the International Neuroimaging Datasharing Initiative. ABIDE collaborated 16 international imaging sites and collected neuroimaging data from 539 individuals suffering from ASD and 573 typical controls (TC). The datasets are composed of structural and resting state functional MRI data along with an extensive array of prototypical information. The major goal of ABIDE is to provide data support to accelerate research of the neural based of ASD (*Di Martino et al.*, 2014).

In the second dissertation topic, an ensemble classification methods is proposed to predict ASD status using a large-scale resting state fMRI data in the Autism Brain Imaging Data Exchange (ABIDE) study data, and lead to a better prediction accuracy than the existing best result.

1.4.3 Electroencephalogram Alcoholism Study

Electroencephalography, EEG, study of alcoholism; see http://kdd.ics.uci.edu/datasets/eeg/eeg.data.html. The objective is to estimate the relationship between alcoholism and brain activity through peak location and magnitude. The study compromises 122 subjects: 77 alcoholic subjects and 45 non-alcoholic controls. For each subject, 64 electrodes were placed on their scalp and EEG was recorded from each electrode at a frequency of 256 Hz (3.9-msec epoch) for 1 second.

In the third dissertation topic, we build a Bayesian model making inference on peaks via spatially adaptive non-stationary Gaussian processes. The method is able to locate peaks of one-dimensional curves and multi-dimensional surfaces. Gaussian process computing techniques are used to achieve efficiency in large scale data settings. We apply the proposed method to analysis of electroencephalogram (EEG) time series in a study of alcoholism. In particular, the proposed method is applied to find the peaks of the EEG time series in the temporal domain and peaks of the signal power in the frequency domain. We also construct a peak-based classifier on alcoholism versus normal controls.

CHAPTER II

Topic 1: Bayesian Gaussian Process Modeling of Large Scale Longitudinal Neuroimaging Data

2.1 Introduction

Developed biomedical technology has made more longitudinal neuroimaging data available, where brain images are collected at several time points. Studying brain activities at repeated occasions is gaining increased attentions in the neuroimaging community because it creates opportunity to further understand the structural and functional development of healthy or pathological brains. The datasets for longitudinal analysis are usually large and complicated, where the response data contain multiple measurements of three-dimensional (3D) brain images for each subject with some of the covariates to be time-varying as well.

2.1.1 Longitudinal Neuroimaging Analysis

Longitudinal neuroimaging analysis is attempt to describe the marginal expectation of the study outcome as a function of given covariates while accounting for the correlation among the repeated imaging measurements. The most common method to fit longitudinal model is the generalized estimating equation (GEE) approach (*Zeger* *and Liang*, 1986), where a mean function and a working correlation structure for the correlation between a subject's repeated measurements are assumed. GEE is only designed for non-high dimensional problem, and its performance is highly limited by the requirement that all covariates have to be time independent, and that the working correlation structure has to be arbitrarily specified (*Sullivan Pepe and Anderson*, 1994; *Fitzmaurice*, 1995). To improve performance of GEE so that covariates are allowed to be time-dependent, *Lai and Small* (2007) proposed the use of generalized method of moment (GMM) for longitudinal data with three types of time-varying covariates and claimed that GMM either outperforms or is comparable with GEE. *Skup et al.* (2012) extended GMM to a multiscale adaptive generalized method of moment (MA-GMM) incorporating spatial information in model estimation based on an iteratively increasing neighborhood concept. Although MA-GMM models between-voxel correlation adaptively to gain computation efficiency, the method fails to work for long range spatial correlations due to the curse of dimensionality. A few literatures tend to explore different modeling strategies for time-varying covariates based on different problem focuses. *Xu et al.* (2008) presented a nonlinear mixed model to study multivariate longitudinal image data that exhibit asymptotic growth trends. Their model requires specification of a parametric function of time for temporal trend. *Li et al.* (2010b) proposed a semi-parametric mixed model, using a nonparametric function based on cubic smoothing splines to model time effects and a parametric function to model other covariate effects. The model uses Bayesian formulation and inference and is applied to longitudinal reproductive hormone dataset. The most recent statistical toolbox conducting longitudinal neuroimaging analysis, called LSTGEE, uses GEE methods but doesn't incorporate spatial dependence (*Li et al.*, 2009).

2.1.2 Machine Learning methods for Neuroimaging Study

Nonparametric methods are more flexible to model spatial dependence of brain images. Principle component techniques are commonly used to extract multiple principle components from spatially correlated imaging data (*Friston et al.*, 1996; *Kerrouche et al.*, 2006). Functional regression models are also an important tool to incorporate complex correlations. *Reiss and Ogden* (2010) used functional principal component regression for image data but without regularization to impose sparsity. *Zhao et al.* (2012) proposed a general wavelet-based Lasso approach in functional linear regression for one dimensional regression coefficient function. *Wang et al.* (2014) proposed a regularized Haar wavelet-based approach that identify brain subregions associated with cognition. For longitudinal analysis, the multilevel functional principal component analysis (MFPCA) by *Di et al.* (2009) used functional principal component as bases and gave description of multilevel functional exposures. *Crainiceanu et al.* (2009) developed generalized multilevel functional linear models (GMFLMs) for association studies between longitudinal outcomes and multilevel functional exposures. They applied their method on imaging data and provided both frequentest and Bayesian inferences. One of the limitations of functional analysis is that it is difficult to find reasonable clinical representation for functional outcomes and regressors. For some cases, complex functional models may have computational bottleneck.

Machine learning tools have also been widely applied for high-dimensional data classification and prediction. Support vector machine methods are among the most popular machine learning techniques due to its efficiency and robustness. Cox and Savoy (2003) used statistical pattern recognition algorithms including LDA and SVM to separate brain activation maps from an fMRI experiment in which participants viewed images of objects. The objects belonged to various categories, both of similar and differing forms. They successfully classified neuroimages into different groups based on what object is viewed when fMRI is conducted. Mitchell et al.(2004) present case studies in which they have successfully trained three classifiers: a Gaussian Naive Bayes classifier, k-nearest neighbor, and linear SVMs, to distinguish cognitive states such as whether the human subject is looking at a picture or a sentence or whether the word the subject is viewing is a word describing food, people, buildings, etc. *Chen and DuBois Bowman* (2011) developed a support vector classifier based on an augmented reproducing kernel function that leverages longitudinal information.

Bayesian spatial modeling approaches have been widely proposed to model the correlation between neighbouring voxels (*Bowman*, 2005). *Marquand et al.* (2010) evaluated the predictive capability of Gaussian process models for two types of quantitative prediction: multivariate regression and probabilistic classification, using whole-brain fMRI volumes from a study investigating subjective responses to thermal pain. They showed that Gaussian process regression models outperform support vector and relevance vector regression. However, the difficulties arise for high dimensional problems since the inversion of the correlation matrix is computationally challenging. To overcome this difficulty, *Banerjee et al.* (2012) introduced the dimension reduction techniques where the original high dimensional space is projected onto a low dimensional subspace with the closest distance to the original matrix.

Based on the literature review above, the majority of current brain study methods are either based on seed voxel analysis or region representative analysis. Due to partial selection of seed voxels or ignoring between regional variation, these methods may lead to substantial information loss and subject to misleading understand in brain functions. There is strong need for a comprehensive method that simultaneously accounts for brain wide spatial and temporal correlations, while adjusting for clinical covariate effects such as age, gender, and medical history. Targeting broad brain area helps reveal potential median to long range changes in brain activities that capable of providing new clues to disease diagnosis, progression, or recovery. We intended to overcome some shortcomings of the existing models by incorporating long range spatial correlations and involving temporal effect through multiple time sessions. We designed a novel Bayesian non-parametric model to predict follow-up neural activities, to identify difference in brain outcomes, and to find potential location and time specific influential covariates. The method incorporates effects of patients' individual characteristics. Both time-varying and time-independent covariates are considered so that our method is not only capable of identifying group specific spatial-temporal influence, i.e. treatment groups, but also able to locate spatial effect of time varying covariates, i.e. task indicator.

2.1.3 Gaussian Processes and Its Properties

Gaussian processes (GPs) are a generalization of multivariate Gaussian distribution to infinitely many dimensions, with a constraint that any finite number has a multivariate normal distribution. GPs are considered to be a powerful tool in various areas, such as non-linear interpolation, supervised and unsupervised machine learning. In statistical modeling, Gaussian process is usually used as a prior probability distribution over functions to make Bayesian inference. Inheriting the good properties of normal distribution, Gaussian process has the advantage of modeling correlation structures so that has been widely used for spatial or temporal models (*Marquand et al.*, 2010; *Gelfand et al.*, 2003). However, the difficulty arises for high dimensional problems for which inversion of the correlation matrix is computationally unfeasible. There has been many explorations in literature to overcome this difficulty. One way to solve this is to use dimension reduction techniques, for example, *Banerjee et al.* (2008) project the original high dimensional space onto a low dimensional subspace that has the closest distance to the original matrix. The other aspect is to add assumptions to involve conditional independence to simplify posterior computation (*Quinonero-Candela and Rasmussen*, 2005). Our research is based on an explicit form of Karhunen-Loéve (K-L) representation of GP using a special kernel function as introduced below.

Suppose $f \sim \mathcal{GP}(\mu, \gamma)$, then there exists a unique orthogonal expansion of $f(\mathbf{v})$ on $\mathcal{B}$, such that

$$f(\mathbf{v}) = \mu(\mathbf{v}) + \sum_{l=1}^{\infty} \theta_l \phi_l(\mathbf{v}), \quad \int\limits_{\mathrm{R}^d} \phi_l(\mathbf{v})\phi_{l'}(\mathbf{v})\mathrm{d}\mathbf{v} = \delta_{ll'}, \qquad \text{and} \quad \theta_l \overset{\text{iid}}{\sim} \mathrm{N}(0, \lambda_l), \quad (2.1)$$

where $\delta_{ll'} = 1$ if $l = l'$ and $\delta_{ll'} = 0$, otherwise. Equation (2.1) is known as the K-L spectral representation of the process $f(\mathbf{v})$. The terms λ_l and functions $\phi_l(\mathbf{v})$ are the eigen values and eigen functions of the covariance kernel $\gamma(\mathbf{v}, \mathbf{v}')$, that is, they are the solutions of the following equation $\gamma(\mathbf{v}, \mathbf{v}') = \sum_{l=0}^{\infty} \lambda_l \phi_l(\mathbf{v})\phi_l(\mathbf{v}')$.

The eigen vectors $\phi_l(\mathbf{v})$'s and eigen values λ_l's have explicit form for certain covariance kernel function with special forms. In one dimensional cases, given the covariance kernel

$$\mathrm{k}_1(x, x'; a, b) = \exp(-ax^2 - b(x - x')^2 - ax'^2),$$

the eigenvalues λ_k's can be calculated as functions of a and b, and the eigenfunctions $\phi_k(\cdot)$'s can be expressed using a, b and normalized hermit polynomial functions. The explicitness can be generalized from one-dimensional to $d-$dimensional kernels

$$\mathrm{k}_d(\mathbf{x}, \mathbf{x}'; a, b) = \prod_{i=1}^{d} \exp(-ax_i^2 - b(x_i - x_i')^2 - ax_i'^2).$$

Detailed derivation is listed in Appendix 2.6.1.

2.2 The Model

Suppose we collect longitudinal imaging data from n subjects at m occasions over the whole three-dimensional (3D) brain $\mathcal{B} \subset \mathbb{R}^d$, where $\mathbb{R}^d$ denotes the $d-$dimensional Euclidean space. For $i = 1, \ldots, n$ and $t = 0, 1, \ldots, m$, denote by $y_{i,t}(\mathbf{v}) \in \mathbb{R}$ the longitudinal imaging outcome at voxel $\mathbf{v} \in \mathcal{B}$ in occasion t for subject i. Denote by $\mathbf{x}_{i,t} = (x_{i,t,1}, \ldots, x_{i,t,p})^{\mathrm{T}} \in \mathbb{R}^p$ a $p-$dimensional vector of time varying covariates for subject i and in time occation t and $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,q})^{\mathrm{T}} \in \mathbb{R}^q$ a $q-$dimensional vector of time independent covariates for subject i.

2.2.1 A Spatial-Temporal Model

For each $\mathbf{v} \in \mathcal{B}$ and each occasion $t = 0, 1, \ldots, m$, we assume the longitudinal imaging outcome

$$y_{i,t}(\mathbf{v}) \quad = \quad \alpha_t(\mathbf{v}) + \sum_{j=1}^{p} x_{i,t,j}\beta_j(\mathbf{v}) + \sum_{k=1}^{q} z_{i,k}\eta_{t,k}(\mathbf{v}) + \epsilon_{i,t}(\mathbf{v}), \tag{2.2}$$

where random error processes $\epsilon_{i,t}(\mathbf{v})$ are mutually independent over subjects, occasions and voxels so that $\epsilon_{i,t}(\mathbf{v}) \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$. In Equation (2.2), the spatially-temporally varying intercept $\alpha_t(\mathbf{v}) \in \mathbb{R}$ is the population-level baseline spatial-temporal effects; the spatially varying coefficient $\beta_j(\mathbf{v})$ is the spatial effects of time varying covariate; the spatially-temporally varying coefficient $\eta_{t,k}(\mathbf{v})$ represents the spatial-temporal effects of time independent covariates.

2.2.2 Prior Specifications

Gaussian processes (GPs) are employed to serve as priors for the spatial-temporal effects in model (2.2). More precisely, let $\mathcal{GP}(\mu, \gamma)$ denote a Gaussian process with mean process $\mu(\mathbf{v})$ and covariance kernel $\gamma(\mathbf{v}, \mathbf{v}')$. For $t = 1, \ldots, m$, $j = 1, \ldots, p$, and

$k = 1, \ldots, q$, we assume

$$\begin{aligned}
[\alpha_t \mid \alpha_{t-1}\ ] &\overset{\text{iid}}{\sim} \mathcal{GP}(\alpha_{t-1}, \tau_\alpha^2 \kappa), \text{and } \alpha_0 \sim \mathcal{GP}(0, \tau_{\alpha,0}^2 \kappa), \\
\beta_j &\overset{\text{iid}}{\sim} \mathcal{GP}[0, \tau_\beta^2 \kappa], \\
[\eta_{t,k} \mid \eta_{t-1,k}\ ] &\overset{\text{iid}}{\sim} \mathcal{GP}[\eta_{t-1,k}, \tau_\eta^2 \kappa], \text{and } \eta_{0,k} \sim \mathcal{GP}(0, \tau_{\eta,0}^2 \kappa), \qquad (2.3)
\end{aligned}$$

where $\kappa(\mathbf{v}, \mathbf{v}')$ is a standardized correlation kernel function for any $\mathbf{v}, \mathbf{v}' \in \mathcal{B}$. Usually, we have $\kappa(\mathbf{v}, \mathbf{v}) = 1$, for $\forall \mathbf{v} \in \mathcal{B}$. In this way, $\boldsymbol{\tau}_{\cdot}^2 = (\tau_\alpha^2, \tau_{\alpha,0}^2, \tau_\beta^2, \tau_\eta^2, \tau_{\eta,0}^2, \tau_\rho^2)$ shows the prior variance of different parameters. For certain κ with special forms that $\kappa(\mathbf{v}, \mathbf{v}) \neq 1$, $\boldsymbol{\tau}^2$ can be considered as a scale parameter that invariant across location, thus $\boldsymbol{\tau}_{\cdot}^2 \kappa(\mathbf{v}, \mathbf{v})$ presents the prior variance of different parameters. Furthermore, the hyperprior parameters for $\boldsymbol{\tau}_{\cdot}^2$ are specified as follows:

$$\begin{aligned}
&\tau_\alpha^2 \sim \mathrm{IG}(a_\alpha, b_\alpha), \qquad \tau_\beta^2 \sim \mathrm{IG}(a_\beta, b_\beta), \qquad \tau_\eta^2 \sim \mathrm{IG}(a_\eta, b_\eta), \\
&\tau_{\alpha,0}^2 \sim \mathrm{IG}(a_\alpha, b_\alpha), \qquad \tau_{\eta,0}^2 \sim \mathrm{IG}(a_{\eta,0}, b_{\eta,0}), \quad \text{and} \quad \sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma), \quad (2.4)
\end{aligned}$$

where $\mathrm{IG}(a, b)$ denotes an inverse-gamma distribution with shape a and scale b. From prior distributions (2.3), we know that $\tau_{\alpha,0}^2$ presents the prior variance of the population-level baseline effects at the initial time occasion, and τ_α^2 presents the prior variance of the population-level baseline effects between time occasions. As a result, $\tau_{\alpha,0}^2$ and τ_α^2 are assumed with different hyperprior distributions in (2.4). The same assumption applies to $\tau_{\eta,0}^2$, the prior variance of the effects of time independent covariates at the initial time occasion, and τ_η^2, the prior variance of the effects of time independent covariates between time occasions.

2.2.3 Model Representation

The standard human brain template usually contains around 200,000 voxels. Thus, posterior inference on the proposed model in Sections 2.2.1 and 2.2.2 for a whole brain analysis involves the extremely challenging ultra-high dimensional GP fitting. To mitigate this problem, we adopt the model representation using the Karhunen-Loéve (K-L) expansion. Suppose the correlation kernel κ in Equation (2.3) has the following expansion:

$$\kappa(\mathbf{v}, \mathbf{v}') = \sum_{l=0}^{\infty} \zeta_l \psi_l(\mathbf{v}) \psi_l(\mathbf{v}'). \tag{2.5}$$

with $\int_{\mathrm{R}^d} \phi_l(\mathbf{v})\phi_{l'}(\mathbf{v})\mathrm{d}\mathbf{v} = \delta_{ll'}$ where $\delta_{ll'} = 1$ if $l = l'$ and $\delta_{ll'} = 0$, otherwise. Then the prior models on α_t, β_j, and $\eta_{t,k}$ can be respectively represented using the K-L expansion as

$$\begin{aligned}
\alpha_t(\mathbf{v}) &= \sum_{l=0}^{\infty} \theta_{\alpha,t,l}\psi_l(\mathbf{v}), \qquad \theta_{\alpha,t,l} \overset{\text{iid}}{\sim} \mathrm{N}\left[\theta_{\alpha,t-1,l}, \tau_\alpha^2 \zeta_l\right], \qquad \theta_{\alpha,0,l} \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau_{\alpha,0}^2 \zeta_l) \\
\beta_j(\mathbf{v}) &= \sum_{l=0}^{\infty} \theta_{\beta,j,l}\psi_l(\mathbf{v}), \qquad \theta_{\beta,j,l} \overset{\text{iid}}{\sim} \mathrm{N}\left[0, \tau_\beta^2 \zeta_l\right] \\
\eta_{t,k}(\mathbf{v}) &= \sum_{l=0}^{\infty} \theta_{\eta,t,k,l}\psi_l(\mathbf{v}), \qquad \theta_{\eta,t,k,l} \overset{\text{iid}}{\sim} \mathrm{N}[\theta_{\eta,t-1,k,l}, \tau_\eta^2 \zeta_l], \qquad \theta_{\eta,0,k,l} \overset{\text{iid}}{\sim} \mathrm{N}[0, \tau_{\eta,0}^2 \zeta_l].
\end{aligned} \tag{2.6}$$

Thus, model (2.2) can be represented as

$$\begin{aligned}
y_{i,t}(\mathbf{v}) &= \sum_{l=0}^{\infty} \theta_{\alpha,t,l}\psi_l(\mathbf{v}) + \sum_{j=1}^{p}\sum_{l=0}^{\infty} \theta_{\beta,j,l}\psi_l(\mathbf{v}) x_{i,t,j} \\
&\quad + \sum_{k=1}^{q}\sum_{l=0}^{L} \theta_{\eta,t,k,l}\psi_l(\mathbf{v}) z_{i,k} + \epsilon_{i,t}(\mathbf{v}).
\end{aligned} \tag{2.7}$$

To create the matrix form of the above linear model, define $\mathbf{0}_d = (\underbrace{0, \ldots, 0}_{d})^{\mathrm{T}}$, $\mathbf{1}_d = (\underbrace{1, \ldots, 1}_{d})^{\mathrm{T}}$, $\mathbf{I}_d = \mathrm{diag}(\mathbf{1}_d)$, $\mathbf{J}_d = \mathbf{1}_d \mathbf{1}_d^{\mathrm{T}}$ and $\mathbf{K}_d = (k_{ij})_{1 \le i,j \le d}$ with $k_{ij} = \min\{i, j\} - 1$. Denote by $\mathbf{y}_i(\mathbf{v}) = [y_{i,0}(\mathbf{v}), \ldots, y_{i,m}(\mathbf{v})]^{\mathrm{T}}$ an $m+1$ vector of the outcome variable. The design matrix is constructed as $\mathbf{D}_i(\mathbf{v}) = [\mathbf{I}_{m+1}, \mathbf{X}_i, \mathbf{z}_i^{\mathrm{T}} \otimes \mathbf{I}_{m+1}]$ of dimension $(m+1) \times Q$, where "$\otimes$" is the kronecker product, $\mathbf{X}_i = (\mathbf{x}_{i,0}, \ldots, \mathbf{x}_{i,m})^{\mathrm{T}}$ of dimension $(m+1) \times p$ with $\mathbf{x}_{i,t} = (x_{i,t,1}, \ldots, x_{i,t,p})^{\mathrm{T}}$, $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,q})^{\mathrm{T}}$, and $Q = m + p + (m+1)q + 1$. The coefficient parameters are constructed as $\mathbf{\Theta}_l = (\boldsymbol{\theta}^{\mathrm{T}}_{\alpha,\cdot,l}, \boldsymbol{\theta}^{\mathrm{T}}_{\beta,\cdot,l}, \boldsymbol{\theta}^{\mathrm{T}}_{\eta,\cdot,\cdot,l})^{\mathrm{T}}$ a $Q \times 1$ vector of coefficients, where $\boldsymbol{\theta}_{\alpha,\cdot,l} = (\theta_{\alpha,0,l}, \ldots, \theta_{\alpha,m,l})^{\mathrm{T}}$, $\boldsymbol{\theta}_{\beta,\cdot,l} = (\theta_{\beta,1,l}, \ldots, \theta_{\beta,p,l})^{\mathrm{T}}$, $\boldsymbol{\theta}_{\eta,\cdot,\cdot,l} = (\boldsymbol{\theta}^{\mathrm{T}}_{\eta,\cdot,1,l}, \ldots, \boldsymbol{\theta}^{\mathrm{T}}_{\eta,\cdot,q,l})^{\mathrm{T}}$ with $\boldsymbol{\theta}_{\eta,\cdot,k,l} = (\theta_{\eta,0,k,l}, \ldots, \theta_{\eta,m,k,l})^{\mathrm{T}}$. From the K-L expansion parameters' prior assumptions in Equation (2.6), it is straightforward to show that

$$\begin{aligned}
\boldsymbol{\theta}_{\alpha,\cdot,l} &\sim \mathrm{N}[\mathbf{0}_{m+1}, \zeta_l(\tau^2_{\alpha,0}\mathbf{J}_{m+1} + \tau^2_{\alpha}\mathbf{K}_{m+1})], \\
\boldsymbol{\theta}_{\beta,\cdot,l} &\sim \mathrm{N}(\mathbf{0}_p, \zeta_l \tau^2_{\beta} \mathbf{I}_p), \\
\boldsymbol{\theta}_{\eta,\cdot,l} &\sim \mathrm{N}[\mathbf{0}_{(m+1)q}, \zeta_l \mathbf{I}_q \otimes (\tau^2_{\eta,0}\mathbf{J}_{m+1} + \tau^2_{\eta}\mathbf{K}_{m+1})]. \qquad (2.8)
\end{aligned}$$

The matrix form of Equation (2.7) is thus given by

$$\mathbf{y}_i(\mathbf{v}) = \mathbf{D}_i(\mathbf{v}) \sum_{l=1}^{\infty} \psi_l(\mathbf{v}) \mathbf{\Theta}_l + \boldsymbol{\epsilon}_i(\mathbf{v}), \qquad (2.9)$$

where $\boldsymbol{\epsilon}_i(\mathbf{v}) \overset{\mathrm{iid}}{\sim} \mathrm{N}(\mathbf{0}_{m+1}, \sigma^2 \mathbf{I}_{m+1})$ with $\boldsymbol{\epsilon}_i(\mathbf{v}) = [\epsilon_{i,0}(\mathbf{v}), \ldots, \epsilon_{i,m}(\mathbf{v})]^T$, and $\mathbf{\Theta}_l \overset{\mathrm{iid}}{\sim} \mathrm{N}(\mathbf{0}_Q, \mathbf{\Sigma}_l)$ with

$$\mathbf{\Sigma}_l = \zeta_l \mathrm{diag}\{(\tau^2_{\alpha,0}\mathbf{J}_{m+1} + \tau^2_{\alpha}\mathbf{K}_{m+1}), \tau^2_{\beta}\mathbf{I}_p, \mathbf{I}_q \otimes (\tau^2_{\eta,0}\mathbf{J}_{m+1} + \tau^2_{\eta}\mathbf{K}_{m+1})\}.$$

2.2.4 Posterior Computation

2.2.4.1 Algorithm I: K-L approximation

For computation purpose, the K-L expansion in Equation (2.5) is usually approximated by finite number of summations,

$$\kappa(\mathbf{v}, \mathbf{v}') \approx \sum_{l=1}^{L} \zeta_l \psi_l(\mathbf{v})\psi_l(\mathbf{v}'). \tag{2.10}$$

where L is sufficient large to control the approximation error. This finite expansion is further applied to the prior models of α_t, β_j, and $\eta_{t,k}$ in Equations (2.6) and then the matrix form regression model in Equation (2.9) so that

$$\mathbf{y}_i(\mathbf{v}) \;\; \approx \;\; \mathbf{D}_i(\mathbf{v}) \sum_{l=1}^{L} \psi_l(\mathbf{v})\mathbf{\Theta}_l + \boldsymbol{\epsilon}_i(\mathbf{v}), \tag{2.11}$$

Let $\mathbf{y} = (\mathbf{y}_1; \ldots; \mathbf{y}_n)$ with $\mathbf{y}_i = (y_i(\mathbf{v}))_{\mathbf{v}\in\mathcal{B}}$, and $\mathbf{\Theta} = (\mathbf{\Theta}_1^T, \ldots, \mathbf{\Theta}_L^T)^T$. Let $\boldsymbol{\tau}_{\cdot}^2 = (\tau_\alpha^2, \tau_{\alpha,0}^2, \tau_\beta^2, \tau_\eta^2, \tau_{\eta,0}^2)$. Denote by $\boldsymbol{\psi}(\mathbf{v}) = (\psi_1(\mathbf{v}), \ldots, \psi_L(\mathbf{v}))^{\mathrm{T}}$ and $\mathbf{\Sigma} = \mathrm{diag}\{\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_L\}$. The full conditional posterior density of $\mathbf{\Theta}$ is given by

$$\begin{aligned} &\pi(\mathbf{\Theta}|\mathbf{y}, \boldsymbol{\tau}_{\cdot}^2, \sigma^2) \propto \pi(\mathbf{y}|\mathbf{\Theta}, \boldsymbol{\tau}_{\cdot}^2, \sigma^2)\pi(\mathbf{\Theta}|\boldsymbol{\tau}_{\cdot}^2) \\ &\quad \propto \;\; \prod_{i=1}^{n}\prod_{\mathbf{v}\in\mathcal{B}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y}_i(\mathbf{v}) - \mathbf{W}_i(\mathbf{v})\mathbf{\Theta}\|^2\right\} \exp\left(-\frac{1}{2}\mathbf{\Theta}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{\Theta}\right) \end{aligned} \tag{2.12}$$

where $\mathbf{W}_i(\mathbf{v}) = \mathbf{D}_i(\mathbf{v})[\boldsymbol{\psi}(\mathbf{v})^{\mathrm{T}} \otimes \mathbf{I}_Q]$. This implies $\mathbf{\Theta}$ given $\mathbf{y}, \boldsymbol{\tau}_{\cdot}^2$ and σ^2 can be drawn from $\mathrm{N}(\boldsymbol{\mu}_{\mathrm{post}}, \mathbf{\Sigma}_{\mathrm{post}})$ with

$$\begin{aligned} \boldsymbol{\mu}_{\mathrm{post}} \;\; &= \;\; \mathbf{\Sigma}_{\mathrm{post}}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{\mathbf{v}\in\mathcal{B}} \mathbf{W}_i(\mathbf{v})^{\mathrm{T}}\mathbf{y}_i(\mathbf{v})\right), \\ \mathbf{\Sigma}_{\mathrm{post}} \;\; &= \;\; \left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{\mathbf{v}\in\mathcal{B}} \mathbf{W}_i^{\mathrm{T}}(\mathbf{v})\mathbf{W}_i(\mathbf{v}) + \mathbf{\Sigma}^{-1}\right)^{-1}. \end{aligned} \tag{2.13}$$

Note that Equation (2.13) evolves calculating the inverse of matrix $\mathbf{\Sigma}$ and matrix $\mathbf{\Sigma}^{-1} + \frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{\mathbf{v}\in\mathcal{B}} \mathbf{W}_i^{\mathrm{T}}(\mathbf{v})\mathbf{W}_i(\mathbf{v})$, both of which are of size $Q \times L$. When $Q \times L$ is small, which may due to that the size of imaging outcomes is moderate, only limited L is required in the K-L expansion in Equation (2.10), and only limited number of time-dependent and time-independent covariates are involved in the model, calculating $\boldsymbol{\mu}_{\text{post}}$ and $\mathbf{\Sigma}_{\text{post}}$ can be direct and fast using Equation (2.13). However, it may arise problem when $Q \times L$ is large. In this case, the following methods are suggested to achieve efficiency.

2.2.4.2 Algorithm II: block updates

Define coefficients $\mathbf{\Theta} = (\mathbf{\Theta}_1^{\mathrm{T}}, \ldots, \mathbf{\Theta}_L^{\mathrm{T}})^{\mathrm{T}}$ with $\mathbf{\Theta}_l = (\boldsymbol{\theta}_{\alpha,\cdot,l}^{\mathrm{T}}, \boldsymbol{\theta}_{\beta,\cdot,l}^{\mathrm{T}}, \boldsymbol{\theta}_{\eta,\cdot,\cdot,l}^{\mathrm{T}})^{\mathrm{T}}$. The full conditional distribution of $\mathbf{\Theta}_l$, given $\mathbf{y}$, $\mathbf{\Theta}_{-l} = \{\mathbf{\Theta}_r, r \in [1, \ldots, L], r \neq l\}$, $\boldsymbol{\tau}^2$, and σ^2 is as follows

$$\begin{aligned}&\pi(\mathbf{\Theta}_l \mid \mathbf{y}, \mathbf{\Theta}_{-l}, \boldsymbol{\tau}^2, \sigma^2) \qquad (2.14)\\ &\propto \prod_{i=1}^{n}\prod_{\mathbf{v}\in\mathcal{B}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y}_i(\mathbf{v}) - \mathbf{D}_i\psi_l(\mathbf{v})\mathbf{\Theta}_l - \mathbf{D}_i\sum_{r\neq l}\psi_r(\mathbf{v})\mathbf{\Theta}_r\|^2\right\}\prod_{l=1}^{L}\exp(-\frac{1}{2}\mathbf{\Theta}_l^T\mathbf{\Sigma}_l^{-1}\mathbf{\Theta}_l).\end{aligned}$$

Thus, $\mathbf{\Theta}_l$, given $\mathbf{y}$, $\mathbf{\Theta}_{-l}$, $\boldsymbol{\tau}^2$, and σ^2 can be drawn from $\mathrm{N}(\boldsymbol{\mu}_{\text{post},l}, \mathbf{\Sigma}_{\text{post},l})$ where

$$\begin{aligned}\boldsymbol{\mu}_{\text{post},l} &= \mathbf{\Sigma}_{post,l} \times \sum_{i=1}^{n}\sum_{\mathbf{v}\in\mathcal{B}} \frac{\psi_l(\mathbf{v})\mathbf{D}_i^{\mathrm{T}}\left(\mathbf{y}_i(\mathbf{v}) - (\sum_{r\neq l}\psi_r(\mathbf{v})\mathbf{D}_i\mathbf{\Theta}_r)\right)}{\sigma^2},\\ \mathbf{\Sigma}_{post,l} &= [\mathbf{\Sigma}_l^{-1} + \sum_{i=1}^{n}\sum_{\mathbf{v}\in\mathcal{B}} \frac{\psi_l^2(\mathbf{v})\mathbf{D}_i^T\mathbf{D}_i}{\sigma^2}]^{-1}, \qquad (2.15)\end{aligned}$$

and $\Sigma_l = \zeta_l \mathrm{diag}\{(\tau_{\alpha,0}^2\mathbf{J}_{m+1} + \tau_\alpha^2\mathbf{K}_{m+1}), \tau_\beta^2\mathbf{I}_p, \mathbf{I}_q \otimes (\tau_{\eta,0}^2\mathbf{J}_{m+1} + \tau_\eta^2\mathbf{K}_{m+1})\}$. Note that comparing posterior distribution of (2.15) with (2.13), the matrix scale that needs inverse operation is reduced to Q from $Q \times L$ so as to achieve computational efficiency.

2.2.4.3 Algorithm III: fast computation

To simply equation (2.15), we take advantage of the orthogonality of ψ_l, where $\int_{\mathbb{R}^3} \psi_l(\mathbf{v})\psi_r(\mathbf{v})d\mathbf{v} = 0$, when $l \neq r$. Given the following two assumptions: 1) $\psi_l(\mathbf{v}) \doteq 0, \forall \mathbf{v} \in \mathcal{B}^c$; 2) there are sufficiently many and equally spaced locations inside $\mathcal{B}$, the following equation applies that $\int_{\mathbf{v}\in\mathbb{R}^3} \psi_l(\mathbf{v})\psi_r(\mathbf{v})d\mathbf{v} \doteq \sum_{\mathbf{v}\in\mathcal{B}} \psi_l(\mathbf{v})\psi_r(\mathbf{v})d\mathbf{v}$. Thus $\sum_{\mathbf{v}\in\mathcal{B}} \psi_l(\mathbf{v})\psi_r(\mathbf{v}) \doteq 0$, $l \neq r$, and posterior distribution (2.15) can be approximated with a short calculation of mean function,

$$\tilde{\boldsymbol{\mu}}_{\text{post},l} \quad = \quad \boldsymbol{\Sigma}_{post,l} \times \sum_{i=1}^{n} \sum_{\mathbf{v}\in\mathcal{B}} \frac{\psi_l(\mathbf{v})\mathbf{D}_i^{\mathrm{T}}\mathbf{y}_i(\mathbf{v})}{\sigma^2}. \tag{2.16}$$

Note that $\boldsymbol{\mu}_{\text{post},l}$ in (2.15) can be approximated by $\tilde{\boldsymbol{\mu}}_{\text{post},l}$ in (2.16) only when assumptions 1) and 2) stand.

2.2.5 An STGPR based Classifier

Suppose one of the time independent variables is categorical, i.e., a group indicator denoted by u_i and taking values of $1, \ldots, G$, where G is the number of groups. Rewrite $\mathbf{z}_i = \{u_i, \mathbf{w}_i\}$ so that $\mathbf{w}_i = \{w_{i,k}, k \in [1, \ldots, q-1]\}$ denote the remaining time-independent variables. Based on the previous regression model, a spatial-temporal Gaussian process classifier is developed in order to gain knowledge of $\Pr(u_i = g), g = 1, \ldots, G$ given observed neuroimaging outcomes $y_{i,t}(\mathbf{v})$, time varying covariates $\mathbf{x}_{i,t}$ and the other time independent covariates $\mathbf{w}_i$ for subject i in time occasion t.

Furthermore, given the priors on u_i specified as

$$\Pr(u_i = g) = \pi_g, \qquad g = 1, \ldots, G,$$

the posterior predictive distribution of u_i is given by

$$\begin{aligned}&\Pr(u_i = g|\mathbf{Y}_i, \mathbf{w}_i, \mathbf{X}_i, \{\mathbf{Y}_l, \mathbf{z}_l, \mathbf{X}_l\}_{l\neq i, l\in\{1,\dots,n\}})\\ \propto\ & \pi_g \int \pi(\mathbf{Y}_i|u_i = g, \mathbf{w}_i, \mathbf{X}_i, \boldsymbol{\Theta}) \prod_{l\neq i} \pi(\mathbf{Y}_l|\mathbf{z}_l, \mathbf{X}_l, \boldsymbol{\Theta})\pi(\boldsymbol{\Theta})d\boldsymbol{\Theta}.\end{aligned}$$

Assume posterior samples of $\boldsymbol{\Theta}$ are calculated as $\boldsymbol{\Theta}^{(1)}, \dots, \boldsymbol{\Theta}^{(N)}$, then an efficient importance sampling can be used to compute leave-one-out cross validation error to examine the predictive performance of the proposed spatial-temporal classifier, where the disease group can be predicted as the one with the largest posterior predictive probability after computing the posterior distribution of u_i as follows:

$$\Pr(u_i = g|\cdot) = \pi_g \sum_{r=1}^{N} \frac{\pi(\mathbf{Y}_i|u_i = g, \mathbf{w}_i, \mathbf{X}_i, \boldsymbol{\Theta}^{(r)})}{\pi(\mathbf{Y}_i|u_i, \mathbf{w}_i, \mathbf{X}_i, \boldsymbol{\Theta}^{(r)})}. \tag{2.17}$$

2.3 Simulation Studies

To show the performance of our model, we conducted Monte Carlo simulation studies, and compared the proposed spatial temporal Gaussian process regression methods (STGPR) with a basic linear regression model (LM), which formulates the same as (2.2) but only to assume mutual independence for $\alpha_t(\mathbf{v})$, $\beta(\mathbf{v})$ and $\eta_t(\mathbf{v})$ over occasions and voxels. Therefore, those parameters are estimated using linear regression method conducted at each individual voxel location and time occasion. For LM prediction, the parameter values at future time points are predicted to be the parameter estimates at the previous time points so that $\hat{\alpha}_m(\mathbf{v}) \equiv \hat{\alpha}_{m-1}(\mathbf{v})$ and $\hat{\eta}_{m,k}(\mathbf{v}) \equiv \hat{\eta}_{m-1,k}(\mathbf{v})$ for $k = 1, \dots, q, \forall \mathbf{v} \in \mathcal{B}$. Then, the predicted outcome variables are calculated by $\hat{f}_{i,m}(\mathbf{v}) = \hat{\alpha}_m(\mathbf{v}) + \sum_{j=1}^{p} x_{i,m,j}\hat{\beta}_j(\mathbf{v}) + \sum_{k=1}^{q} z_{i,k}\hat{\eta}_{m,k}(\mathbf{v})$ at future time m. Point and standard error estimates for all parameters in LM are provided by least square estimators.

We provided two simulation settings with different imaging size to demonstrate

Algorithm II and Algorithm III accordingly. Scenario 1, utilizing Algorithm II, is designed over spatial data of size 60×60 and equally spread on two dimensional space of $[-1.8, 1.8]^2$. Scenario 2, utilizing Algorithm III, is designed over 100×100 equally spaced points on $[-1.8, 1.8]^2$, where 1)$[-1.8, 1.8]^2$ constrains the effective region with nonzero eigen functions and 2)the fine grids of 100×100 enables the orthogonal approximation. Both simulation scenarios contains 20 samples each measured at four sequential time occasions.

For both scenarios, two time-dependent covariates are generated by $x_{i,1}(t) = 1.5 \times t + \epsilon_{i,1,t}$ and $x_{i,2}(t) = 2.5 \times t + \epsilon_{i,2,t}$ respectively, where $\epsilon_{i,j,t} \sim_{i,t} \mathcal{N}(0, 1), j = 1, \ldots, p$. These time-dependent covariates are considered as continuous variables and are standardized over subjects and over time before being incorporated into the model. In addition, two time-independent covariates, $z_{i,t,1}$ and $z_{i,t,2}$ are sampled independently from Bernoulli distribution with success probability to be 0.5 and 0.1. The coefficients, $\alpha_t, \beta_j, \eta_{t,k}$ are generated based on equation (2.3) and $\kappa(\mathbf{v}, \mathbf{v}') = e^{-\frac{d(\mathbf{v},\mathbf{v}')}{s}}$ with $d(\mathbf{v}, \mathbf{v}')$ calculating the l^2 distance between in $\mathcal{B}$. The value of hyperprior parameters in the simulation study can be found in Table (2.1).

Table 2.1: The hyperprior parameters used in simulation study to generate baseline function, coefficients, and random noise

α			
$\tau^2_{\alpha_0} = 1$	$s_{\alpha_0} = 1$	$\tau^2_{\alpha} = 0.1$	$s_{\alpha} = 1$
η_1			
$\tau^2_{\eta_{0,1}} = 0.5$	$s_{\eta_{0,1}} = 1$	$\tau^2_{\eta_1} = 0.1$	$s_{\eta_1} = 1$
η_2			
$\tau^2_{\eta_{0,2}} = 0.5$	$s_{\eta_{0,2}} = 1$	$\tau^2_{\eta_2} = 0.1$	$s_{\eta_2} = 1$
β			
$\tau^2_{\beta_1} = 0.5$	$s_{\beta_1} = 1$	$\tau^2_{\beta_2} = 0.1$	$s_{\beta_2} = 1$
σ^2			
$\sigma^2 = 1$			

In order to show both estimation and prediction performance of STGPR compared to LM, we divide the simulated data into two parts: the first three time occasions for model fitting and the fourth time occasion for prediction. Here, two statistical measures are designed to evaluate the performance of LM and STGPR in regards to both estimation and prediction of model parameters and imaging outcomes. The relative L_1 Loss (RL_1) averages over space the absolute values of the percentage bias between model calculates and truth. The gradients' relative L_1 Loss (GRL_1) measures a spatial average of the absolute percentage difference between the first derivatives of model calculates and the truth. For time varying parameters or outcomes, RL_1 Loss and GRL_1 Loss perform an average over both time and space. Assume $g(\mathbf{v})$ to represent any functions with respect to voxels $\mathbf{v} \in \mathcal{B}$ and $\hat{g}(\mathbf{v})$ to be the model estimates. Let B represent the total number of voxels. For time-independent functions like $\beta_j(\mathbf{v}), \forall j$, the above statistical measures are formulated as:

$$\begin{aligned} \mathrm{RL}_1(g) &= \frac{1}{B}\sum_{\mathbf{v}\in\mathcal{B}} \left|\frac{\hat{g}(\mathbf{v}) - g(\mathbf{v})}{g(\mathbf{v})}\right| \\ \mathrm{GRL}_1(g) &= \frac{1}{dB}\sum_{\mathbf{v}\in\mathcal{B}}\sum_{s=1}^{d} \left|\frac{\partial\hat{g}(\mathbf{v})/\partial v_s - \partial g(\mathbf{v})/\partial v_s}{\partial g(\mathbf{v})/\partial v_s}\right| \end{aligned}$$

where $\partial g/\partial v_s$ represent the s-th partial derivative for function g. For time-dependent functions like $\alpha_t(\mathbf{v}), \eta_{t,k}(\mathbf{v}), \forall k$ and $y_t(\mathbf{v})$, the above statistical measures are formulated as follows:

$$\begin{aligned} \mathrm{RL}_1(g) &= \frac{1}{B(m-1)}\sum_{\mathbf{v}\in\mathcal{B}}\sum_{t=1}^{m-1} \left|\frac{\hat{g}_t(\mathbf{v}) - g_t(\mathbf{v})}{g_t(\mathbf{v})}\right| \\ \mathrm{GRL}_1(g) &= \frac{1}{dB(m-1)}\sum_{\mathbf{v}\in\mathcal{B}}\sum_{t=1}^{m-1}\sum_{s=1}^{d} \left|\frac{\partial\hat{g}_t(\mathbf{v})/\partial v_s - \partial g_t(\mathbf{v})/\partial v_s}{\partial g_t(\mathbf{v})/\partial v_s}\right| \end{aligned}$$

The statistical measures for prediction of time-dependent variables at $t = m$ are:

$$\begin{aligned} \mathrm{RL}_1(g_{\text{pred}}) &= \frac{1}{B}\sum_{\mathbf{v}\in\mathcal{B}} \left|\frac{\hat{g}_m(\mathbf{v}) - g_m(\mathbf{v})}{g_m(\mathbf{v})}\right| \\ \mathrm{GRL}_1(g_{\text{pred}}) &= \frac{1}{dB}\sum_{\mathbf{v}\in\mathcal{B}}\sum_{s=1}^{d} \left|\frac{\partial\hat{g}_m(\mathbf{v})/\partial v_s - \partial g_m(\mathbf{v})/\partial v_s}{\partial g_m(\mathbf{v})/\partial v_s}\right| \end{aligned}$$

Table (2.2) showed the simulation results for Algorithm II and Algorithm III over

100 repeated trials. We can see that LM and STGPR have comparable estimation performance, however STGPR provides a smaller and stabler prediction of parameters and image outcomes.

Table 2.2: Simulation results comparing LM and STGPR in regards to estimation and prediction of both coefficients and outcome variables through RL$_1$ Loss(relative L_1 Loss) and GRL$_1$ (gradients' relative L_1 Loss) averaging over space and necessarily over time. For STGPR, Algorithm II is used on simulated data of size 60×60 and Algorithm III is used for simulated data of size 100×100. The values in the parenthesis are standard error of the quantities of interest.

Algorithm II	**Size**(60×60)			
	LM		STGPR	
	RL$_1$	GRL$_1$(×10)	RL$_1$	GRL$_1$(×10)
α	1.06(0.07)	244.66(434.68)	0.76(0.02)	9.52(1.10)
β_1	1.31(0.25)	10.91(7.68)	1.09(0.04)	1.13(0.39)
β_2	0.73(0.07)	12.00(2.53)	0.77(0.02)	1.99(0.65)
η_1	0.94(0.06)	7.52(6.55)	1.17(0.02)	0.97(0.54)
η_2	0.61(0.05)	3.15(0.69)	0.42(0.01)	0.45(0.01)
f	0.37(0.01)	3.92(1.89)	1.01(0.01)	1.84(0.56)
α_{pred}	5.15(0.36)	9.92(16.71)	4.85(0.06)	1.12(0.41)
$\eta_{1,\text{pred}}$	3.12(0.16)	24.40(20.02)	3.22(0.03)	5.60(4.08)
$\eta_{2,\text{pred}}$	5.11(0.96)	15.35(36.78)	4.31(0.10)	5.53(0.98)
y_{pred}	2.57(0.21)	7.39(26.32)	2.93(0.01)	1.51(0.76)
Algorithm III	**Size**(100×100)			
	LM		STGPR	
	RL$_1$	GRL$_1$(×10)	RL$_1$	GRL$_1$(×10)
α	1.20(0.15)	10.79(17.30)	0.60(0.02)	1.21(3.30)
β_1	1.02(0.08)	159.41(1078.06)	0.52(0.03)	1.28(4.83)
β_2	0.65(0.06)	6.70(7.88)	0.25(0.01)	0.60(0.41)
η_1	1.38(0.23)	2.53 (0.87)	0.41(0.03)	0.32(0.05)
η_2	0.59(0.05)	2.37(2.07)	0.43(0.01)	0.27(0.08)
f	1.17(0.64)	2.11(0.53)	0.81(0.15)	0.98(0.94)
α_{pred}	3.13(0.15)	3.37 (3.32)	2.73(0.05)	0.72(0.54)
$\eta_{1,\text{pred}}$	4.91(0.44)	15.60(7.61)	4.82(0.08)	7.36(25.39)
$\eta_{2,\text{pred}}$	4.06(0.73)	6.30(9.66)	3.87(0.13)	2.03(0.53)
y_{pred}	2.16(0.09)	1.56(0.68)	2.19(0.02)	0.69(0.57)

Based on the above simulation results, we can conclude that both LM and STGPR provide reasonable parameter and outcome estimation and prediction. The LM model is more influenced by random noise, while STGPR model provides stabler estimation with smaller standard errors of RL$_1$ and GRL$_4$ than LM model. The STGPR model results in a much smaller GRL$_4$ than the LM model for all the parameters and outcome variables, indicating that the smoothness of the estimated curves by the proposed method are much closer to the truth. As a result, the proposed STGPR model provides stable outcome prediction and reliable estimates of the covariates' spatial and temporal effects.

2.4 ADNI Data Analysis

In order to show the usefulness of our method, we illustrate the model using PET images of patients from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (http://www.loni.ucla.edu/ADNI/). The goal of this national multi-center project is to develop biomarkers of Alzheimer's Disease (AD) in elderly subjects. For more details about the ADNI, see Mueller et al. (2005). Participants are classified as having mild cognitive impairment (MCI), as Alzheimer's disease (AD) patients, or as typical controls (TC). PET scans are obtained for each participants at baseline (screening) and 6 months and 12 months. The processing steps for the PET scans include co-registration, averaging, and standardizing image and voxel size into a standard $160 \times 160 \times 96$ image grid and spatial smoothing. In addition, we perform a spatial normalization to a standard $91 \times 109 \times 91$ MNI space (*Tzourio-Mazoyer et al.*, 2002).

In this study, we apply STGPR to brain images of 219 subjects collected at baseline, 6 months, and 12 months sequentially and include a subset of available covariates collected by ADNI as potential regressive predictors. For the time-independent covariates, we include gender and diagnostic status (AD, MCI, or TC) and investigate their corresponding spatial and temporal influence on imaging responses. For time dependent covariates, we use age and weight and considered their spatial-varying effects on the response images. The diagnostic status is treated as the group indicator. We also performed classification analysis based on the results from STGPR. The brain images are divided into 116 regions based on automated anatomical labeling (*Tzourio-Mazoyer et al.*, 2002). In Alzheimer's disease, the hippocampus regions are known by researchers to be one of the first brain area to suffer damage in that good prediction on brain response at these locations can contribute to studies of AD initiation and development. We particularly choose the two hippocampus regions to test the performance of STGPR.

Five-fold cross validation is conducted to compare STGPR and LM for their predictive performances. When dividing individual data into five folds, it is guaranteed that at least 4 AD individuals are in each fold to facilitate balanced estimation and prediction. The predictive results are listed in Table (2.3). STGPR shows smaller or comparative predicted mean squared error (PMSE) than LM, specifically for AD group or for MCI group on brain region of Hippocampus left.

Table 2.3: A summary of results for ADNI data analysis. MSE(mean squared error) and PMSE(predicted mean squared error) are provided based on five-fold cross validation conducting on both Hippocampus left and right regions and for all subjects, AD group, and MCI group respectively.

Hippocampus_L

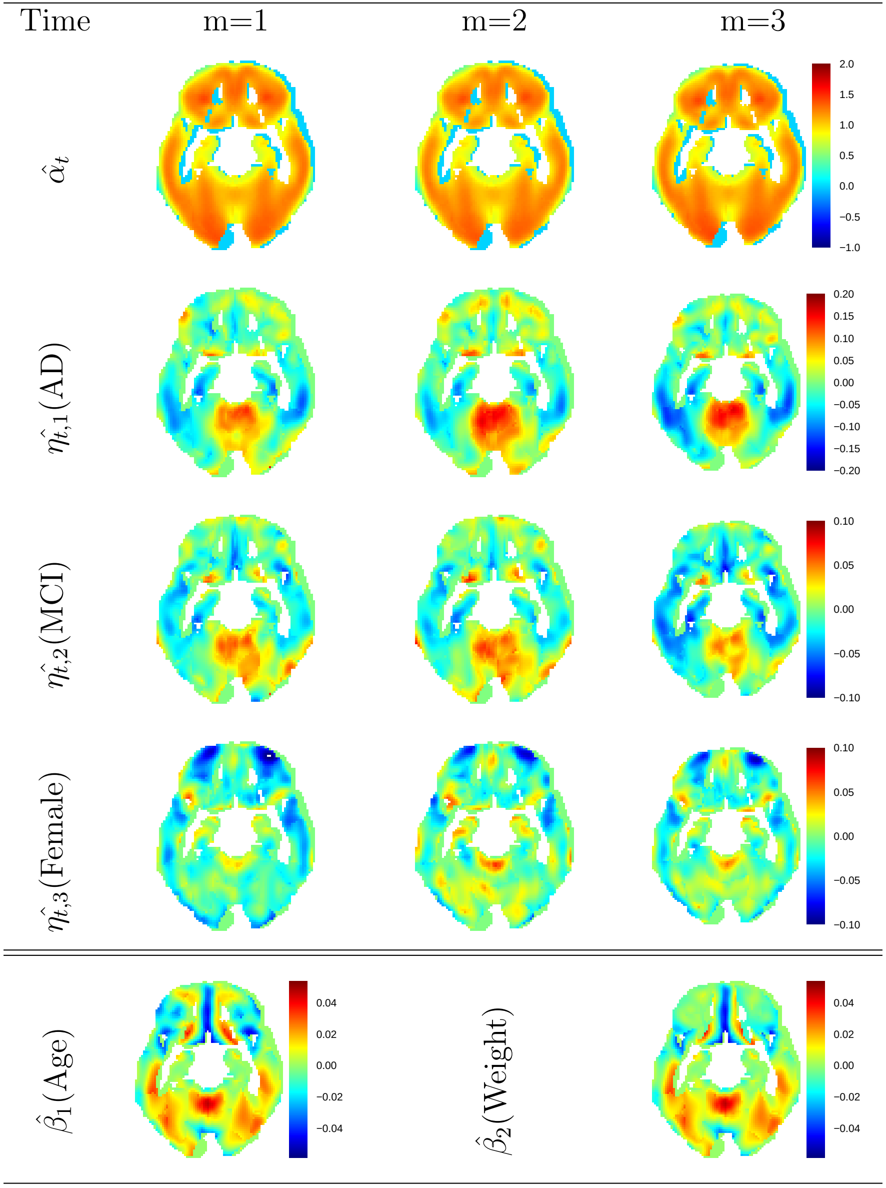
	All Subjects		AD		MCI	
	MSE	PMSE	MSE	PMSE	MSE	PMSE
LM	0.01109	0.01813	0.01210	0.008612	0.01029	0.02464
STGPR	0.01112	0.01812	0.01215	0.008603	0.01036	0.02464

Hippocampus_R

	All Subjects		AD		MCI	
	MSE	PMSE	MSE	PMSE	MSE	PMSE
LM	0.01032	0.01727	0.01042	0.00567	0.00972	0.02434
STGPR	0.01034	0.01726	0.01047	0.00565	0.00975	0.02434

We also apply the STGPR on the slice of neuroimages that covers most areas of Hippocampus regions. The parameter estimation results of the use area are shown in Figure (2.1). The spatial maps of covariates' effects at different time occasions estimated by STGPR are provided. $\hat{\alpha}_t$'s in the first row provided the estimated population-level baseline spatial-temporal effects; $\hat{\eta}_{t,1}$'s (or $\hat{\eta}_{t,2}$'s) showed the estimated mean spatial and temporal difference in brain responses between AD (or MCI) and NORM subjects when other covariates are the same. $\eta_{t,3}$ demonstrated the estimated mean spatial and temporal difference in brain response between female and male subjects given all other covariates remain the same. $\hat{\beta}_1$ or $\hat{\beta}_2$ provided a spatial map of brain response changes given unit change in age or weight when all other covariates stays the same.

Figure 2.1: Maps of coefficients' effects estimated by proposed STGPR for ADNI data. Plots include population baseline effects, $\hat{\alpha}_t$, spatial effects for time-varying covariates, $\hat{\beta}_j$, and spatial-temporal effects for time-independent covariates, $\hat{\eta}_{t,j}$



2.5 Discussion

We develop a spatial-temporal Gaussian process regression model to localize brain activities and characterize their changes over time. The proposed model takes into account spatial correlation of imaging outcome and the temporal correlation between successive longitudinal imaging outcomes over space. The proposed model is also capable of making inference on the spatial effects of time varying covariates and the spatial-temporal effects of time independent covariates on the brain activities. Also, it can be used to make prediction on brain activities. We develop fast posterior computation algorithms for model fitting, which are computationally efficient and feasible for high-dimensional neuroimaging data. It takes a regular computer with i5 cpu and 8G memory less than 1 hours to complete the analysis of the largest brain region which contains about 5100 voxels. An interesting consideration for future research would be to conduct brain-wise analysis by specifying a between region convariance structure and make prediction on the disease status using longitudinal imaging data of the entire brain area.

There are rather limited study in the brain imaging literature for longitudinal neuroimaging models, especially when spatial correlation structures are considered over broad brain areas. It is also unique for our study to consider spatial and temporal structures in covariates' effects. Furthermore, the fast computational algorithms proposed provide an option for extremely large scale brain imaging analysis. The above listed are contribution of this work to large scale longitudinal neuroimaging study.

2.6 Appendix

2.6.1 Explicit Forms of Eigen Values and Eigen Vectors

In one-dimensional cases, for covariance kernel

$$\mathrm{k}_1(x, x') = \exp(-ax^2 - b(x - x')^2 - ax'^2),$$

define

$$\begin{aligned} c &= \sqrt{a^2 + 2ab}, \\ A &= a + b + c, \\ B &= b/A. \end{aligned}$$

For $k = 0, 1, \ldots, n$, the kth eigenvalues λ_k and eigenfunctions $\phi_k(\cdot)$ are respectively given by

$$\begin{aligned} \lambda_k &= \sqrt{\frac{\pi}{A}} B^k, \\ \phi_k(x) &= (\sqrt{2c})^{1/2} \exp(-cx^2) \widetilde{H}_k(\sqrt{2c}x), \\ \mathrm{k}_1(x, x') &= \sum_{k=0}^{\infty} \lambda_k \phi_k(x) \phi_k(x'), \end{aligned}$$

where $\widetilde{H}_k(\cdot)$ is the kth order normalized hermit polynomial, which is defined by

$$\widetilde{H}_k(x) = (2^n n! \sqrt{\pi})^{-1/2} \exp(x^2) \frac{d^n}{dx^n} \exp(-x^2).$$

For $d-$dimensional case ($d \geq 2$), we have covariance kernel

$$\mathrm{k}_d(\mathbf{x}, \mathbf{x}'; a, b) = \prod_{i=1}^{d} \exp(-ax_i^2 - b(x_i - x_i')^2 - ax_i'^2).$$

The connection between $\mathrm{k}_d(\mathbf{x}, \mathbf{x}')$ and $\mathrm{k}_1(x_i, x_i')$ can be given by

$$\begin{aligned}
\mathrm{k}_d(\mathbf{x}, \mathbf{x}'; a, b) &= \prod_{i=1}^{d} \mathrm{k}_1(x_i, x_i') = \prod_{i=1}^{d} \sum_{k=0}^{\infty} \lambda_k \phi_k(x_i) \phi_k(x_i') \\
&= (\lambda_0 \phi_0(x_1) \phi_0(x_1') + \ldots)(\lambda_0 \phi_0(x_2) \phi_0(x_2') + \ldots)(\lambda_0 \phi_0(x_3) \phi_0(x_3') + \ldots) \ldots \\
&= \sum_{k=0}^{\infty} \sum_{m_1 + m_2 + \cdots m_d = k} \prod_{i=1}^{d} \lambda_{m_i} \phi_{m_i}(x_i) \phi_{m_i}(x_i') \\
&= \sum_{k=0}^{\infty} \sum_{m_1 + m_2 + \cdots m_d = k} \prod_{i=1}^{d} \sqrt{\frac{\pi}{A}} B^{m_i} \phi_{m_i}(x_i) \phi_{m_i}(x_i') \\
&= \sum_{k=0}^{\infty} \sum_{m_1 + m_2 + \cdots m_d = k} \prod_{i=1}^{d} \sqrt{\frac{\pi}{A}} B^{m_i} \times \prod_{i=1}^{d} \phi_{m_i}(x_i) \phi_{m_i}(x_i') \\
&= \sum_{k=0}^{\infty} \left(\sqrt{\frac{\pi}{A}} \right)^d B^k \sum_{m_1 + m_2 + \cdots m_d = k} \prod_{i=1}^{d} \phi_{m_i}(x_i) \phi_{m_i}(x_i') \\
&= \sum_{l=0}^{\infty} \lambda_l \psi_l(\mathbf{x}) \psi_l(\mathbf{x}')
\end{aligned}$$

The eigenvalues and eigenfunctions of $d-$dimensional kernel are given by

$$\begin{aligned}
\lambda_l &= \left(\sqrt{\frac{\pi}{A}} \right)^d B^k, \qquad \binom{k-1+d}{d} \leq l \leq \binom{k+d}{d} - 1 \\
\widetilde{\lambda}_k &= \sum_{\binom{k-1+d}{d} \leq l \leq \binom{k+d}{d} - 1} \lambda_l = \binom{k+d-1}{d-1} \left(\sqrt{\frac{\pi}{A}} \right)^d B^k \\
\sum_{k=0}^{\infty} \widetilde{\lambda}_k &= \frac{1}{(1-B)^d} \left(\sqrt{\frac{\pi}{A}} \right)^d \\
\frac{\sum_{k=0}^{m} \widetilde{\lambda}_k}{\sum_{k=0}^{\infty} \widetilde{\lambda}_k} &= (1-B)^d \sum_{k=0}^{m} \binom{k+d-1}{d-1} B^k
\end{aligned}$$

2.6.2 Full Conditional Posterior Density of Hyperparameters

The full conditional posterior density of $\tau_{\alpha,0}^2$ is given by

$$
\begin{aligned}
&\pi(\tau_{\alpha,0}^2|\mathbf{y}, \boldsymbol{\Theta}, \boldsymbol{\tau}_{-\alpha,0}^2) \\
&\quad \propto \; \pi(\boldsymbol{\Theta}|\tau_{\cdot}^2)\pi(\tau_{\alpha,0}^2) \\
&\quad \propto \; \prod_{l=1}^{L} |\zeta_l(\tau_{\alpha,0}^2\mathbf{J}_{m+1} + \tau_{\alpha}^2\mathbf{K}_{m+1})|^{-1/2} \exp(-\frac{1}{2}\sum_{l=1}^{L} \boldsymbol{\theta}_{\alpha,\cdot,l}^T[\zeta_l(\tau_{\alpha,0}^2\mathbf{J}_{m+1} + \tau_{\alpha}^2\mathbf{K}_{m+1})]^{-1}\boldsymbol{\theta}_{\alpha,\cdot,l}) \\
&\qquad \times p_{inv-gamma}(\tau_{\alpha,0}^2; a_{\alpha,0}, b_{\alpha,0})
\end{aligned}
$$

The full conditional posterior density of τ_{α}^2 is given by

$$
\begin{aligned}
&\pi(\tau_{\alpha}^2|\mathbf{y}, \boldsymbol{\Theta}, \boldsymbol{\tau}_{-\alpha}^2) \\
&\quad \propto \; \pi(\boldsymbol{\Theta}|\tau_{\cdot}^2)\pi(\tau_{\alpha}^2) \\
&\quad \propto \; \prod_{l=1}^{L} |\zeta_l(\tau_{\alpha,0}^2\mathbf{J}_{m+1} + \tau_{\alpha}^2\mathbf{K}_{m+1})|^{-1/2} \exp(-\frac{1}{2}\sum_{l=1}^{L} \boldsymbol{\theta}_{\alpha,\cdot,l}^T[\zeta_l(\tau_{\alpha,0}^2\mathbf{J}_{m+1} + \tau_{\alpha}^2\mathbf{K}_{m+1})]^{-1}\boldsymbol{\theta}_{\alpha,\cdot,l}) \\
&\qquad \times p_{inv-gamma}(\tau_{\alpha}^2; a_{\alpha}, b_{\alpha})
\end{aligned}
$$

The full conditional posterior density of τ_{β}^2 is given by

$$
\begin{aligned}
&\pi(\tau_{\beta}^2|\mathbf{y}, \boldsymbol{\Theta}, \boldsymbol{\tau}_{-\beta}^2) \\
&\quad \propto \; \pi(\boldsymbol{\Theta}|\tau_{\cdot}^2)\pi(\tau_{\beta}^2) \\
&\quad \propto \; \prod_{l=1}^{L} |\zeta_l\tau_{\beta}^2\mathbf{I}_p|^{-1/2} \exp(-\frac{1}{2}\sum_{l=1}^{L} \boldsymbol{\theta}_{\beta,\cdot,l}^T[\zeta_l\tau_{\beta}^2\mathbf{I}_p]^{-1}\boldsymbol{\theta}_{\beta,\cdot,l}) \times p_{inv-gamma}(\tau_{\beta}^2; a_{\beta}, b_{\beta}) \\
&\quad \propto \; p_{inv-gamma}(\tau_{\beta}^2; a_{\beta} + \frac{Lp}{2}, b_{\beta} + \frac{1}{2}\sum_{l=1}^{L}\frac{\boldsymbol{\theta}_{\beta,\cdot,l}^T\boldsymbol{\theta}_{\beta,\cdot,l}}{\zeta_l})
\end{aligned}
$$

The full conditional posterior density of $\tau^2_{\eta_0}$ is given by

$$\begin{aligned}
&\pi(\tau^2_{\eta,0}|\mathbf{y},\mathbf{\Theta},\boldsymbol{\tau}^2_{-\eta,0}) \\
&\quad\propto \pi(\mathbf{\Theta}|\tau^2_{\cdot})\pi(\tau^2_{\eta,0}) \\
&\quad\propto \prod_{l=1}^{L}|\zeta_l(\tau^2_{\eta,0}\mathbf{J}_{m+1}+\tau^2_{\eta}\mathbf{K}_{m+1})|^{-q/2} \\
&\qquad \times\exp(-\frac{1}{2}\sum_{l=1}^{L}\sum_{k=1}^{q}\boldsymbol{\theta}^T_{\eta,\cdot,k,l}[\zeta_l(\tau^2_{\eta,0}\mathbf{J}_{m+1}+\tau^2_{\eta}\mathbf{K}_{m+1})]^{-1}\boldsymbol{\theta}_{\eta,\cdot,k,l}) \\
&\qquad \times p_{inv-gamma}(\tau^2_{\eta,0};a_{\eta,0},b_{\eta,0})
\end{aligned}$$

The full conditional posterior density of τ^2_{η} is given by

$$\begin{aligned}
&\pi(\tau^2_{\eta}|\mathbf{y},\mathbf{\Theta},\boldsymbol{\tau}^2_{-\eta}) \\
&\quad\propto \pi(\mathbf{\Theta}|\tau^2_{\cdot})\pi(\tau^2_{\eta}) \\
&\quad\propto \prod_{l=1}^{L}|\zeta_l(\tau^2_{\eta,0}\mathbf{J}_{m+1}+\tau^2_{\eta}\mathbf{K}_{m+1})|^{-q/2} \\
&\qquad \times\exp(-\frac{1}{2}\sum_{l=1}^{L}\sum_{k=1}^{q}\boldsymbol{\theta}^T_{\eta,\cdot,k,l}[\zeta_l(\tau^2_{\eta,0}\mathbf{J}_{m+1}+\tau^2_{\eta}\mathbf{K}_{m+1})]^{-1}\boldsymbol{\theta}_{\eta,\cdot,k,l}) \\
&\qquad \times p_{inv-gamma}(\tau^2_{\eta};a_{\eta},b_{\eta})
\end{aligned}$$

The full conditional posterior density of σ^2 is given by

$$\begin{aligned}
&\pi(\sigma^2|\mathbf{y},\mathbf{\Theta}) \\
&\quad\propto \pi(\mathbf{y}|\mathbf{\Theta},\sigma^2)\pi(\sigma^2) \\
&\quad\propto \prod_{i=1}^{n}\prod_{\mathbf{v}\in\mathcal{B}}(\sigma^2)^{-(m+1)/2}\exp(-\frac{1}{2\sigma^2}\|\mathbf{y}_i(\mathbf{v})-\mathbf{W}_i(\mathbf{v})\mathbf{\Theta}\|^2\times p_{inv-gamma}(\sigma^2;a_{\sigma^2},b_{\sigma^2}) \\
&\quad\propto p_{inv-gamma}(\sigma^2;a_{\sigma^2}+\frac{nd(m+1)}{2},b_{\sigma^2}+\frac{\sum_{i=1}^{n}\sum_{\mathbf{v}\in\mathcal{B}}\|\mathbf{y}_i(\mathbf{v})-\mathbf{W}_i(\mathbf{v})\mathbf{\Theta}\|^2}{2})
\end{aligned}$$

CHAPTER III

Topic 2: Ensemble Classification Methods For Feature Combination of Large Scale Neuroimaging Data

3.1 Introduction

For many neurological diseases, to predict disease status and forecast disease progression is of great clinical importance and has very influential therapeutic meanings. For example, autism spectrum disorders (ASDs), which refers to a syndrome of social communication deficits and repetitive behaviors or narrow interest, usually appears during patients' infancy or childhood. Researches have shown that early identification of ASD and a subsequent early and intensive behavioral intervention help to improve patients' cognitive function and decrease symptom severity (*Fein et al.*, 2013; *Rogers and Vismara*, 2008). Given that the ASDs have an estimated prevalence of 1:68 in the USA, research into the early identification of ASDs represents a public health priority (*Developmental et al.*, 2014). Another example could be Alzheimer's disease (AD), which leads to the death of nerve cells and tissue loss throughout the brain. When patients start to experience dementia, which usually appears to be the indicative symptoms of AD, the disease usually has already caused irreversible brain damage. Early diagnosis of AD will provide an opportunity for early medical intervention so as to effectively slow down disease progression, reduce lost of brain functions, and to the best lead to successful therapeutic treatment (**?**).

Due to the fact, disease diagnosis is one of the major tasks of statistical neuroimaging studies. To achieve this goal, imaging data, which provides detailed information about brain structure and function information, are utilized as predictive biomarkers or quantitative traits to imply disease categories of interest. There are many statistical challenges for categorizing patients' disease groups using neuroimaging data. First, neuroimaging data are often extremely high throughput in the order of hundreds of thousands to millions voxels but with very few samples, which raises both practical and theoretical challenges of high dimensionality problems for model fitting (*Fan and Fan*, 2008). Second, from a neuroscience point of view, a human brain is considered a complex system that multiple brain regions are anatomically connected and functionally interact with each other. As so, neuroimaging data, considered as predictors, usually include spatial correlation. Complex spatial correlation especially under high dimensional scenarios brings challenge to both model building and computation. Third, patients could have brain images taken at various time occasions. The temporal change of multiple consequent measurements may also relate to disease diagnosis and progression. However, in most cases, only limited number of longitudinal records are available in neuroimaging study, for example, MRI scan that taken at baseline, after 6 months, after 12 months, etc. This requires a model to capture temporal change from only a few sequential inputs while accommodating spatial variation. Fourth, patients' demographic information and other relative covariates may also reflect disease status. How to include these factors into disease prediction and correctly model their underlying correlation with neuroimaging biomarkers also brings challenge to build the classification model.

Many research have been conducted for disease diagnosis using neuroimaging data. Both statistical models and machine learning methods have been shown success in disease classification. In the following, section (3.1.1) provides a detailed review of existing methods and their limitations. Section (3.1.2) provides an introduction of the support vector machine classifier which the proposed method is based on. Section (3.1.3) reviews existing ensemble classification methods, where the proposed method is chosen from.

3.1.1 Disease Prediction Using Neuroimaging Data

Many efforts have been made for disease diagnosis utilizing a combination of means including clinical measures, cognitive test, neuroimaging study, genetic biomarkers, and etc. Neuroimaging scans, such as structural MRI (sMRI), functional MRI (fMRI), positron emission tomography (PET), directly record brain activities and have been used broadly of diagnosis of a number of medical disorders and illnesses.

Many statistical methods and multiple machine learning techniques have been employed to leverage neuroimaging data for disease discrimination study. The main idea is to identify disease related features to distinguish between disease groups and typical controls (TC). However, one of the main challenges in imaging classification is the high dimensionality of the feature space. Many learning based methods directly use low-level features related to anatomical brain structures for discrimination among disease groups. These features, relatively low dimension, are extracted from neuroimaging data and are designed to capture disease related variations. Some example of the low-level features includes ventricles size, hippocampus shape, cortical thickness, brain volume, and etc. Based on these low-level features, typical classifiers can be used to distinguish disease groups and TC. Among the most popular are linear discriminant analysis (LDA), neural network (NN) and support vector machines (SVM) (*French et al.*, 1997; *Savio et al.*, 2009; *Magnin et al.*, 2009). LDA aims to find the mapping that reduces the input dimensionality, while preserving the most class discriminatory information. *Adeli-Mosabbeb et al.* (2015) proposed a LDA method that is robust to sample-outliers and feature-noises and tested on two brain neurodegenerative disease, particularly for Parkinsons disease and AD. *Belmokhtar and Benamrane* (2012) proposed a multiclass classification method based on binary SVM to distinguish between patients with AD, patients with mild cognitive impairment (MCI) and elderly control subjects. *Chen and DuBois Bowman* (2011) developed a novel support vector classifier for longitudinal high dimensional data that leverages the additional longitudinal neuroimaging information to achieve better classification performance.

Recent studies have shown that feature fusion from multiple imaging modalities can enhance the diagnostic performance, i.e. gray matter tissue volume from MRI, mean signal intensities from PET (*Hinrichs et al.*, 2011; *Suk et al.*, 2015). *Kohannim et al.* (2010) concatenated low-level features from different modalities into a vector and trained a SVM classifier. *Hinrichs et al.* (2011) and *Zhang et al.* (2011) utilized a multi-kernel SVM to combine information from multimodal data. *Shi et al.* (2014) took into account the association among low-level features extracted from neuroimaging data and devised a coupled feature representation by utilizing intra-coupled and inter-coupled interaction relationship. Regarding multi-modal feature fusion, they proposed a coupled boosting algorithm that analyzes the pairwise coupled-diversity correlation between modalities. There are also a growing number of publications that use deep learning to solve neuroimaging classification problem, i.e. manifold learning with restricted Boltzmann machine (*Brosch et al.*, 2013), a multiclass deep learning framework with modified k-sparse autoencoder (*Bhatkoti and Paul*, 2016), and a meta analysis of different state of the art approaches (*Plis et al.*, 2014).

Although the above listed researches presented the effectiveness of their methods for neuroimaging classification, the main limitation is that they considered only simple low-level features. However, some latent information inherent in the original neuroimaging data also provide helpful to improve model performance. Utilizing neuroimaging data from the whole brain incorporate most information so that lead to improved performance for disease classification. With increase computation power, there are more methods that take advantage of high throughput voxel level neuroimaging data to perform disease diagnosis. *Guo et al.* (2015) developed a supervised dimension reduction framework, called spatially weighted principal component analysis, for high-dimensional imaging classification. *Filippone et al.* (2012) proposed a multi-modality multinomial logit model using Gaussian process as priors to predict disease status based on whole-brain neuroimaging data and analyze the relative informativeness of different image modalities and brain regions. Their method was used for discrimination of three Parkinsonian neurological disorders from one another and healthy controls and showed a promising predictive performance. *Hosseini-Asl et al.* (2016) proposed to predict the AD with a deep 3D convolutional neural network, which can learn generic features capturing AD biomarkers and adapt to different domain datasets. The 3D-CNN is built upon a 3D convolutional autoencoder, which is pre-trained to capture anatomical shape variations in structural brain MRI scans. Fully connected upper layers of the 3D-CNN are then fine-tuned for each task-specific AD classification. *Zhu et al.* (2014) proposed a novel matrix similarity based loss function for joint regression and classification in AD diagnosis.

In this study, we investigate ensemble methods inspired by boosting strategy, which is proposed by *Demiriz et al.* (2002), for its use in predicting Autism disease. First multiple SVM are trained as basic classifiers to predict disease status based on different sources, containing demographic and clinical information, neuroimaging data from separate brain regions and at multiple time occasion, and etc. After classification results are collected based on different sources, a linear programming boost approach is conducted to find optimal combination of basic classifiers so as to achieve maximized classification performance. We also provide insights of when combination methods can be expected to work and how the benefit of complementary features can be exploited most efficiently.

3.1.2 Support Vector Machine Classifiers

In machine learning, support vector machine classifiers are supervised learning methods that classify outcome variable into different categories without using probabilistic models. It has been shown to have good performance to solve challenging classification problem in a wide range of application domains. The original idea of SVM is proposed by *Vapnik* (1963) to choose the linear separating hyperplane that maximize the margin between the hyperplane and the closest examples. Linear SVM is then generalized to nonlinear by applying kernel trick (*Aronszajn*, 1944; *Boser et al.*, 1992). The current popular and widely used SVM which allows some examples to violate the constraints by introducing soft margins was introduced by *Vapnik* (1995). The new loss function in this case includes penalization term of training set errors.

Typically an SVM approach requires the solution of a quadratic programming (QP) or a linear programming (LP) problem depending on whether to use L_2 norm or L_1 norm to measure the margin (*Wu and Zhou*, 2005; *Kecman and Hadzic*, 2000). The general QP algorithms via quasi-Newton methods or primal-dual interior-point methods can only handle problem of small sample size, i.e. thousands of points. The LP algorithms based on simplex or interior points can solve problems of moderate size, i.e. hundreds of thousands of data points (*Bennett and Campbell*, 2000). The scale of SVM using QP and LP is also limited by the possible computer memories since these algorithms need operations on the original data matrix or the kernel matrix. There are several approaches to solve SVM for larger datasets include techniques where kernel components are discarded after evaluated (*Frie et al.*, 1998), chunking and decomposition methods where subset of data is used (*Joachims*, 1998; *Collobert and Bengio*, 2001; *Platt et al.*, 1999; *Keerthi et al.*, 2001), and etc. In neuroimaging study, since the collected data samples are usually limited, QP algorithms are commonly chosen to solve SVM.

The most common technique in practice to do multiclass classification with SVM is to build K one-versus-rest classifiers and to choose the class which classifies the data with greatest margin. Another strategy is to build a set of one-versus-one classifiers, and to choose the class that is selected by the most classifiers, which involves building $K(K-1)/2$ classifiers. The time used by the one-versus-one strategy may actually decrease compared to one-versus-all strategy, since the training dataset for each classifier is much smaller. Some authors also proposed methods that consider all classes in one optimization formulation (*Weston and Watkins*, 1998; *Crammer and Singer*, 2002). Among all, one-against-one approach based on binary SVM classifier has been shown competitive and suitable to use in practice in terms of accuracy and computational cost (*Pal*, 2008; *Hsu and Lin*, 2002). One of the most popular SVM librariy LIBSVM choose to implement the one-versus-one methods for SVM (*Chang and Lin*, 2011).

3.1.3 Ensemble Classification Methods

Ensemble methods are learning algorithms that combine results from multiple classifications to improve performance. The initial classifications can be made by one or several traditional classifiers, e.g. Decision Trees, Neutral Networks, SVM, and etc., and can be trained on data from different physical domains, or from different types of analysis, or even from repeated samples from the same data. The results from these individual classifiers can be weak with low predictive power, while the combination improves the final classification power. There are various types of ensemble rules that has shown improved performance by combining the basic set of classifiers, especially for complicated datasets (*Duin and Tax*, 2000; *Kittler et al.*, 1998). The popular combination ideas include majority vote that selects the most frequently assigned class label, averaging methods of additive equally weighted classifiers, product methods of multiplicative equally weighted classifiers, and optimal linear combination methods to optimize jointly over a linear combination of classifiers, etc. Some commonly used ensemble methods include Bayesian averaging, that the ensemble consists of all single classifiers weighted by its posterior probability, Bootstrap aggregating (*Breiman*, 1996), and boosting. Different classifier ensemble methods have been widely applied to different areas including hand writing classification (*Xu et al.*, 1992), image classification (*Gehler and Nowozin*, 2009) and etc.

Boosting is a machine-learning method that boost the accuracy of weak or base classifiers by assigning each classifier an additive weight and evaluate their aggregate response. The weak classifiers, when considered individually may have low predictive power, usually show improved results when combine. The underlying premise is that if the weak classifiers' errors are uncorrelated, their combination gives a better approximation of the underlying signal. The early and most well known boosting algorithm for binary classification is AdaBoost algorithm (*Freund and Schapire*, 1995). Since it was first introduced, various versions of the Adaboost algorithm have been used in a variety of applications and have proven to be very competitive with each other in terms of prediction accuracy (*Friedman et al.*, 2000; *Efron et al.*, 2004; *Grove and Schuurmans*, 1998), among which Linear Programming boosting (LPboost), first proposed by *Demiriz et al.* (2002), fit the boosting approach into a linear optimization framework with a soft margin bias. At first, LPBoost is considered intractable for large dataset because of the scale of optimization (*Breiman*, 1999). After a column generation based simplex method was proposed (*Demiriz et al.*, 2002) as efficient solution to solve LPBoost, it has been widely used in a variate of applications, especially when used as online learning algorithms in computer vision, (*Saffari et al.*, 2010).

The idea of LPBoost is to separate the feature space into two regions, where each region contains either positively or negatively labeled examples. The weights of different basic classifiers are achieved by maximizing the margin between the positive and negative regions. Compared to early boosting methods such as AdaBoost, LPBoost adjust all weights of existing basic classifiers every time when a new classifier is added so that allows faster convergence. In contrast to gradient boosting algorithms, which may only converge in the limit, LPBoost converges in a finite number of iterations to a globally optimal solution satisfying well-defined optimal conditions. The optimal solutions of LPBoost are very sparse in contrast with gradient based methods. In regards to computational cost, an iteration of LPBoost is slightly more expensive than an iteration of AdaBoost, but on the other hand LPBoost needs far fewer iterations than AdaBoost to converge.

3.2 Methods

Suppose we collect neuroimaging data from n subjects over the whole three-dimensional (3D) brain $\mathcal{B} \subset \mathbb{R}^d$, where $\mathbb{R}^d$ denotes the $d-$dimensional Euclidean space, and $d = 3$. Assume the whole brain is divided into r separate regions such that $\mathcal{B} = \cup_{k=1}^r \mathcal{B}_k$, $\mathcal{B}_k \subset \mathbb{R}^d$, and $\mathcal{B}_k \cap \mathcal{B}_{k'} = \emptyset$ when $1 \leq k \neq k' \leq r$. Let s_k representing the number of voxels in brain region k, and let $s = \sum_{k=1}^r s_k$ representing the total number of voxels in the whole brain. For $i = 1, \ldots, n$, denote by $y_i(\mathbf{v}) \in \mathbb{R}$ the imaging outcome at voxel $\mathbf{v} \in \mathcal{B}$ for subject i. Denote $\mathbf{y}_i^{(k)} = \{y_i(\mathbf{v}), \mathbf{v} \in \mathcal{B}_k\}$, a s_k dimensional vector of imaging signals at brain region k for subject i. Denote $\mathbf{y}_i = \bigcup_{k=1}^r \mathbf{y}_i^{(k)} = \{y_i(\mathbf{v}), \mathbf{v} \in \mathcal{B}\}$, a s dimensional vector of imaging signals covering all brain regions for subject i. For $i = 1, \ldots, n$, let g_i represent the group indicator taking values of $1, \ldots, c$, where c is the number of groups. For example, in the ABIDE dataset, $c = 2$ represent the typical control (TC) and disease group.

Other than the above notation of the observed data records, we choose to use capital letters to represent the corresponding random variables and random vectors. Denote $Y(\mathbf{v}) \in \mathbb{R}$ a random variable representing the brain signal at voxel $\mathbf{v}$, $\mathbf{Y}^{(k)} =$

$\{Y_(\mathbf{v}), \mathbf{v} \in \mathcal{B}_k\}$ a random vector representing brain signals at brain region $\mathcal{B}_k$, and $\mathbf{Y} = \{\mathbf{Y}^{(k)}, k = 1, \ldots, r\}$ a random vector representing brain signals of the whole brain. Denote G a random variable representing the disease group. The interest of this study is to determine $\Pr(G = g|\mathbf{Y})$ with $g = 1, \ldots, c$, the probability of a patient belong to group g given observed neuroimaging data $\mathbf{Y}$ of all brain regions. Since the scale of $\mathbf{Y}$, which is s, can be very large, we proposed to build separate probabilistic model for each brain region and use LPBoost methods to combine the predictive results together. As so, we first build r basic classifier as

$$f_k : \mathbf{Y}^{(k)} \longrightarrow G \in \{1, \ldots, c\},$$

mapping the s_k dimensional random vector $\mathbf{Y}^{(k)}$ to random variable G indicating group labels. Each of the basic classifiers is to be used to predict group label using neuroimaging data from brain region k. Second, we use an ensembled classifier to combine the classification results from all brain region basic classifiers. The ensembled classifier takes advantage of each basic classifier so as to improve the final classification performance. In the following, we will first introduce support vector classifier and its extensions which are good candidates for basic classifiers. Then we will introduce linear programming boosting, an ensemble classification method that optimize combination of all basic classifiers.

Note that when longitudinal neuroimaging data is provided such that there are multiple observation of brain images at different time occasions. For example $\mathbf{y}_{i,t}$, and $t = 0, \ldots m$. Separate basic classifiers are suggested to be constructed based on brain images at each brain regions and every single time occasions, and LPBoost is then used to combine predictive results from all basic classifiers. Moreover, if additional feature besides neuroimaging data are collected, e.g. time-varying covariates $\mathbf{x}_{i,t} \in \mathbb{R}^p$, time-independent covariates $\mathbf{w}_i \in \mathbb{R}^q$, or genetic information $\mathbf{a}_i \in \mathbb{R}^l$. Then the method can be generalized to model $\Pr(G_i = g|\mathbf{y}_{i,t}, \mathbf{x}_{i,t}, \mathbf{w}_i, \mathbf{a}_i)$. Separate basic classifiers should be constructed based on time-varying covariates, time-independent covariantes or genetic information independently and then combined together for ensemble performance. However, for simplicity, the presentation below only shows model formulation based on single time neuroimaging data.

3.2.1 Binary Support Vector Classifiers

In this section, we demonstrate the use of support vector machine and its role as a basic classifier to be used in our model. First we consider a binary classification problem with two categories. Without loss of generality, we create a mapping function to map random variable $G \in \{1, 2\}$ and random variable $O \in \{-1, 1\}$, so that

$$o(G) = \begin{cases} -1 & G = 1 \\ 1 & G = 2 \end{cases} \tag{3.1}$$

$O = o(G)$ also represents a binary state associated with disease statues. According to the mapping, $O = -1$ represent group of typical controls (TC), and $O = 1$ represent group of patients with disease (PD). Binary variable O is the output of SVM classifiers.

The idea of building a SVM classifier can be explained by constructing a separating function $h(\cdot) = \mathbf{w} \cdot \Phi(\cdot) + b$ that separate neuroimaging data satisfying $h(\cdot) \leq 0$ into group TC, and the ones satisfying the opposite, $h(\cdot) \geq 0$, into group PD. Feature vectors that are on hyperplanes $|h(\cdot)| = 1$ are called supporting vectors. When data from the two groups are completely separable, supporting vectors are points from the two groups that are with the shortest distance to the separating hyperplane $h(\cdot) = 0$. Thus the separating function can be obtained by maximizing the margin, which is defined as the distance between two hyperplanes, i.e. $h(\cdot) = -1$ for group TC and $h(\cdot) = 1$ for group PD. For cases when data are not completely separable, a 'soft margin' is introduced that allows some neuroimaging data to be misclassified. In this case, the separating function aims to maximize the margin and at the same time controlling the distance between misclassified data points from hyperplane $|h(\cdot)| = 1$ of the correct side. $\Phi(\cdot)$ in the separating function is a feature function that can either be linear or nonlinear which determines the linearity of the SVM classifer. When $\Phi(\cdot)$ is nonlinear, it performs nonlinear transformation from the original input space to a transformed feature space where the separating hyperplane is built. Through $\Phi(\cdot)$, the SVM classifier is able to map input data into a higher dimension, where separation of the input data becomes achievable. Instead of defining $\Phi(\cdot)$ directly, the nonlinear transformation is often defined through kernel functions, characterized by $K(\mathbf{X}, \mathbf{X}')$, the dot product of $\Phi(\mathbf{X})^{\mathrm{T}}\Phi(\mathbf{X}')$, where $\mathbf{X}$ represents the input random variables. The kernel trick allows the transformed space to be generalized to some unspecified high dimensions (*Boser et al.*, 1992). Some commonly used kernels include homogeneous polynomial, inhomogeneous polynomial, hyperbolic tangent, and etc. In this study, Gaussian radial basis function kernel, or RBF kernel, is used where

$$K(\mathbf{X}, \mathbf{X}') = \exp(-\gamma||\mathbf{X} - \mathbf{X}'||), \gamma > 0.$$

The RBF kenel is one of the most commonly used kernel in SVM and has shown good performance (*Chang et al.*, 2010).

In this problem, we formulate the SVM classifiers for region k as

$$\begin{aligned} f_k(\mathbf{Y}^{(k)}) &= o^{-1}[\text{sign}\{h_k(\mathbf{Y}^{(k)})\}] && (3.2) \\ h_k(\mathbf{Y}^{(k)}) &= \mathbf{w}_k \cdot \Phi_k(\mathbf{Y}^{(k)}) + b_k, && (3.3) \end{aligned}$$

where $h_k(\cdot) = 0$ defines the separating hyperplane for region k, and $\mathbf{w}_k$ and b_k are the corresponding parameters. Given observations of n subjects, including neuroimaging data and the corresponding group labels denoted by $\mathbf{y}_i^{(k)}$ and $o_i \in \{-1, 1\}$, $i = 1, \ldots, n$, the SVM parameters can be solved by minimizing the following optimization function:

$$\min \qquad \mathcal{P}_k(\mathbf{w}_k, b_k) = \frac{1}{2}||\mathbf{w}_k||^2 + C_k \sum_i \mathrm{H}[o_i h_k(\mathbf{Y}_i^{(k)})]. \tag{3.4}$$

Geometrically, $2/||\mathbf{w}_k||^2$ quantifies the distance between the two hyperplanes $|h_k(\cdot)| = 1$. The Hinge loss $\mathrm{H}(\cdot)$, which is defined as $\mathrm{H}(z) = \max(0, 1 - z)$, quantifies the misclassified distance. The constraint constant C_k is the tuning parameter regarding the tolerance level of misclassification.

To solve (3.4), we introduce slack variable $\xi_i^{(k)} = \max[0, 1 - o_i\{\mathbf{w}_k \cdot \Phi_k(\mathbf{Y}_i^{(k)}\} + b_k)]$, and $\xi_i^{(k)}$ is the smallest nonnegative number satisfying $o_i(\mathbf{w}_k \cdot \Phi_k(\mathbf{Y}_i^{(k)}) + b_k) \leq 1 - \xi_i^{(k)}$. Problem (3.4) can be rewritten as the following constrained optimization problem:

$$\begin{aligned} \min \qquad & \mathcal{P}_k(\mathbf{w}_k, b_k, \boldsymbol{\xi}^{(k)}) = \frac{1}{2}||\mathbf{w}_k||^2 + C_k \sum_{i=1}^{n} \xi_i^{(k)} \\ \text{subject to} \qquad & o_i(\mathbf{w}_k \cdot \Phi_k(\mathbf{Y}_i^{(k)}) + b_k) \geq 1 - \xi_i^{(k)} \\ & \xi_i^{(k)} \geq 0, \forall i \end{aligned} \tag{3.5}$$

Consider (3.5) as a primal problem. In order to solve the primal problem, the Lagrangian dual of (3.5) is introduced to obtain a simplified optimization:

$$\begin{aligned} \max \qquad & \mathcal{D}_k(\alpha_1^{(k)}, \ldots, \alpha_n^{(k)}) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i^{(k)} o_i \Phi_k(\mathbf{Y}_i^{(k)})^T \Phi_k(\mathbf{Y}_j^{(k)}) o_j \alpha_j^{(k)} + \sum_{i=1}^{n} \alpha_i^{(k)} \\ \text{subject to} \qquad & \sum_{i=1}^{n} \alpha_i^{(k)} o_i = 0, \text{and } 0 \leq \alpha_i^{(k)} \leq C_k, \forall i \end{aligned} \tag{3.6}$$

The dual problem (3.6) is an optimization with a convex quadratic objective and only linear constraints, which can be efficiently solved by quadratic programming methods.

Assume $\hat{\boldsymbol{\alpha}}^{(k)} = (\hat{\alpha}_1^{(k)}, \ldots, \hat{\alpha}_n^{(k)})$ is a solution of the dual problem (3.6), the solution of the primal problem (3.5) is then recovered as:

$$\begin{aligned}\hat{\mathbf{w}}_k &= \sum_{i=1}^{n} \hat{\alpha}_i^{(k)} o_i \Phi_k(\mathbf{Y}_i^{(k)}) && (3.7)\\ \hat{b}_k &= \arg\min_b \sum_{i=1}^{n} [1 - o_i(\hat{\mathbf{w}}_k \cdot \Phi_k(\mathbf{Y}_i^{(k)}) + b]_+ && (3.8)\end{aligned}$$

where $\hat{\mathbf{w}}_k$ is the direction of the hyperplane and $\hat{b}_k$ is the intercept. When there is new data with brain signals $\mathbf{Y}_{\text{new}}^{(k)}$ to be classified, the linear discriminant function is first calculated

$$h_k(\mathbf{Y}_{\text{new}}^{(k)}) = \sum_{i=1}^{n} o_i \hat{\alpha}_i^{(k)} \Phi_k(\mathbf{Y}_i^{(k)})^{\mathrm{T}} \Phi_k(\mathbf{Y}_{\text{new}}^{(k)}) + \hat{b}_k, \qquad (3.9)$$

and the class label is assigned based on $o(\text{sign}(h_k(\mathbf{Y}_{\text{new}}^{(k)}))$.

SVM have become a popular technique to handle classification problems, though there are some drawbacks. First SVMs scale with the data size due to the quadratic optimization algorithm and the kernel transformation. This cause challange when solving problem with large dataset. Second, the correct choice of kernel parameters is crucial for obtaining good results, which practically means that an extensive search must be conducted on the parameter space before arriving a trustful results, which often complicates the task. Recent algorithms for finding the SVM classifier include sub-gradient descent and coordinate descent. Both techniques have proven to offer significant advantages over the traditional approach when dealing with large, sparse datasets. Sub-gradient methods are especially efficient when there are many training examples, and coordinate descent when the dimension of the feature space is high. In this study, since the number of samples for neural imaging study often appears to be relative small, traditional quadratic programming is used to solve SVM. The implementation is based on LIBSVM package (*Chang and Lin*, 2011).

3.2.2 Binary LPBoosting through Column Generation

Continuing to solve the original classification problem, we target to classify a patent's disease group based on the neuroimaging data collected from r brain regions. Assume that r independent basic SVM classifiers are already built corresponding to different brain regions. Denote the separating functions for the k-th SVM classifier as $h_k(\cdot)$ with parameters $\hat{\mathbf{w}}_k$ and $\hat{b}_k$, $k = 1, \ldots, r$. A boosting ensemble classifier is a linear combination of basic classifiers via optimized weights. In this problem, the final classifier can be written as

$$f_{\text{ensemble}}(\{\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(r)}\}) \quad = \quad o(\text{sign} \sum_{k=1}^{r} \beta_k h_k(\mathbf{Y}^{(k)})),$$

where $\mathbf{Y}^{(k)}, k = 1, \ldots, r$ is the random variables representing neuroimaging data from r regions. Denote by $\mathbf{H}$ a matrix with elements $H_{ik} = h_k(\mathbf{y}_i^{(k)})$, representing the label given by the k-th classifier on the training sample i, $k = 1, \ldots, r, i = 1, \ldots, n$. In order to solve the linear weights, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)$, a linear program can be formulated with the same soft margin cost used in SVM (*Demiriz et al.*, 2002):

$$\max_{\boldsymbol{\beta}, \rho, \boldsymbol{\varepsilon}} \quad \mathcal{P}(\boldsymbol{\beta}, \rho, \boldsymbol{\varepsilon}) = \rho - D \sum_{i=1}^{n} \varepsilon_i, \tag{3.10}$$

$$\text{subject to} \quad \begin{cases} \forall i, \ o_i \cdot \sum_{k=1}^{r} \beta_k H_{ik} \geq \rho - \varepsilon_i \\ \forall i, \ \varepsilon_i \geq 0 \\ \forall k, \beta_k \geq 0, \text{and} \sum_{k=1}^{r} \beta_k = 1. \end{cases} \tag{3.11}$$

Here, $\varepsilon_i \geq 0$ are the slack variables, showing the scale of how wrongly sample i is misclassified. ρ represents margin, quantifying the distinction between two classes. The constant factor $D > 0$ is the tradeoff parameter between misclassification error and margin maximization.

A dual problem of (3.10) is written as

$$\min_{\mu,\gamma} \quad \mathcal{D}(\gamma) = \gamma, \tag{3.12}$$

$$\text{subject to} \quad \begin{cases} \forall i,\ \sum_{k=1}^{r} \mu_k o_i H_{ik} \leq \gamma \\ \forall k, 0 \leq \mu_k \leq D, \text{and} \sum_{k=1}^{r} \mu_k = 1 \end{cases} \tag{3.13}$$

Comparing the primal solution that gives the weightings of the weak learners, the dual solution provide the distributions over examples. In (3.13), we can consider μ_i as misclassification cost to each train sample points such that the μ_i sum to 1. The complementarity of the primal solution (3.10) and the dual solution (3.12) can be expressed as equality of the primal and dual objectives:

$$\rho - D\sum_{i=1}^{n} \varepsilon_i = \gamma, \tag{3.14}$$

and the weights of basic classifiers, β_k's, in the primal problem and the weights of samples, μ_i's,in the dual problem have the following property:

$$\mu_i(o_i \sum_{k=1}^{r} H_{i,k}\beta_k + \varepsilon_i - \rho) = 0, i = 1, \ldots, n \tag{3.15}$$

$$\beta_k(\sum_{i=1}^{n} \mu_i o_i H_{i,k} - \gamma) = 0, k = 1, \ldots, r. \tag{3.16}$$

The dual constraint $\sum_{i=1}^{n} \mu_i o_i H_{ik}$ in (3.13) scores each weak classifier $h_k(\cdot)$ by calculating the weighted sum of correctly classified points minus the weighted sum of the incorrectly classified points. For a given $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, the set of best weak classifiers will have a score of γ. By minimizing γ, the dual problem finds the optimal $\boldsymbol{\mu}$ that maximize the primal objective. From the complementary slackness conditions (3.15) and (3.16), only the training samples that are misclassified and that are on the margin have positive misclassification cost μ_i, and only the basic classifier with scores equal to γ can associate to positive weights β_j in the primal space.

When the number of basic classifiers are large, traditional linear programming methods, i.e. simplex method, and interior point method, are considered intractable. The classic technique of column generation avoids taking into consideration all the variables explicitly, so that it is shown to solve LPBoost efficiently. Especially when there are huge number of variables compared to the number of constraints, most variables are set to be zero in the optimal solution, and only a subset of variables need to be considered. Column generation leverages this idea to generate only the variables which have the potential to improve the objective function, instead of enumerating all possibilities. The problem is first formulated as a restricted master problem (RMP). This RMP has as few variables as possible, and new variables are brought in only when it is associated with a negative reduced cost under current dual variables. The variable with the most negative reduced cost is added to the RMP and the master program is re-solved. Resolving the master program will generate a new set of dual values and this process is repeated until no negative reduced cost variables are identified. When all the restrictions are satisfied, the solution can be concluded to be optimal. The detailed algorithm of column generation is as follows:

```
ALGORITHM: LPBoost

    INPUT:

            $D, \mathbf{H}$

    INITIALIZATION:

            $j \leftarrow 0$, no weak hypotheses are included

            $\boldsymbol{\beta} \leftarrow 0$, all coefficients are 0

            $\gamma \leftarrow 0$

            $\boldsymbol{\mu} \leftarrow (\frac{1}{n}, \ldots, \frac{1}{n})$, corresponding optimal dual
```

REPEAT:

$j \leftarrow j + 1$

Find $l(j) = \arg \max_{k \in (1,\ldots,r)} \sum_{i=1}^{n} \mu_i o_i H_{ik}$

If $\sum_{i=1}^{n} u_i o_i H_{il(r)} \leq \gamma$, then $j \leftarrow j - 1$, break

Otherwise, solve RMP for new costs $\boldsymbol{\mu}$ and γ:

$$\begin{aligned} \text{Goal:} \quad & \arg\min \gamma \\ \text{Constraint:} \quad & \sum_{i=1}^{n} u_i o_i H_{il(a)} \leq \gamma, a = 1, \ldots, j \\ & 0 \leq u_i \leq D, i = 1, \ldots, n \end{aligned} \tag{3.17}$$

OUTPUT:

$\boldsymbol{\beta} \leftarrow$ Lagrangian multipliers from last LP

$f = \sum_{a=1}^{j} \beta_a h_{l(a)}$

Practically we found $D = \frac{1}{n\nu} \in (\frac{1}{n}, 1)$ preferable because of the interpretability of the parameter. By picking ν appropriately we can force the minimum number of support vectors, which is the number of points misclassified plus the points on the margin.

3.2.3 Multiclass SVM and LPBoost

We extend our method to deal with situation when there are more than two groups to be classified, e.g. patients with three Alzheimer status including typical control(TC), MCI, and AD. Denote c the number of classes and $c \geq 3$. Without loss of generality, multiclass SVM is used as basic classifiers. In practice, the multiclass SVM decomposes multiclass problem into a series of binary classification such that the standard two class SVM can be directly applied. Two representative schemes are one-versus-rest (*Vapnik*, 1998) and one-versus-one (*Kreßel*, 1999) approaches. The one versus rest method builds c classifiers, and chooses the class which classifies the test data with greatest margin. The one versus one methods build a classifier for every two of the $\mathcal{C}$ groups and choose the class that is selected by the most classifiers. While this involves building $c(c-1)/2$ classifiers, the time for training all classifiers may actually decrease, since the training dataset for each classifier is much smaller compared to one versus the rest method (*Weston et al.*, 1999; *Demiriz et al.*, 2002; *Gehler and Nowozin*, 2009). Although the multiclass problem can also be addressed in one single optimization process that combines multiple binary class optimization into one single objective function and simultaneously achieves classification of multiple classes, a larger computational complexity is required in this case due to the size of the resulting quadratic programming problem. In this study, the one-versus-one strategy is used as provided in the implementation of LIBSVM package.

After multiclass basic SVM classifiers are trained, the results are provided to multiclass LPBoost to find optimal ensemble solution. To continue with previous notation, we assume that there are c classes, and r basic classifiers from different regions. Denote $\mathbf{h}_k(\cdot)$ the k'th basic classifier. Instead of having a real valued output for binary classification, $\mathbf{h}_k(\cdot)$ takes feature input $\mathbf{Y}^{(k)}$ from region k and map into a $\mathbb{R}^c$ dimensional space. Denote $\mathbf{H}_{ik}$ the outcome of the k'th basic classifier for subject i, so that $\mathbf{H}_{ik}$ is a vector of length c. Denote H_{ikc} the c'th elements in $\mathbf{H}_{ik}$. Note that $\mathbf{H}_{ik}$ are trained by basic classifiers in the first step.

Following the idea of using multi-class SVM, multiclass LPBoost can also be decomposed into binary LPBoost problem. One versus rest LPBoost builds linear combination function $f^{(a)}_{\text{ensemble}}$ to distinguish in class a versus not in class a. After a total number of c binary LPBoost models are built, the predicted label is choose to be the class with the highest $f^{(a)}_{\text{ensemble}}$ score. One versus one LPboost builds a number of $c(c-1)$ binary LPboost and take the predicted label as the class with the majority vote.

Other than relying on binary LPboost, there are two variation of multiclass LP-Boost methods based on over all optimization but having different feature weights' assumption. The first method, LP-β, uses a single vector $\boldsymbol{\beta}$ of length r as ensemble weights for all classes jointly. Alternatively, the second method, LP-B, defines a weight matrix $B \in \mathbb{R}^{r\times c}$ so that each class preserves its own ensemble weight vector over c features spaces.

In LP-β, the ensemble rules are:

$$f_{\text{ensemble}}(\{\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(r)}\}) = \arg \max_{g\in\{1,\ldots,c\}} \sum_{k=1}^{r} \beta_k \mathbf{h}_k(\mathbf{Y}^{(k)}), \tag{3.18}$$

The mixing coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ are learned by the following multiclass extension of LPBoost:

$$\max_{\boldsymbol{\beta},\rho,\boldsymbol{\xi}} \quad \mathcal{P}(\boldsymbol{\beta}, \rho, \boldsymbol{\xi}) = \rho - \frac{1}{\nu n}\sum_{i=1}^{n} \varepsilon_i, \tag{3.19}$$

$$\text{subject to} \quad \begin{cases} \forall i,\ \sum_{k=1}^{r} \beta_k H_{iko_i} - \arg\max\limits_{o_j \neq o_i} \sum_{k=1}^{r} \beta_k H_{ikg_j} \geq \rho - \varepsilon_i, \\ \forall i,\ \varepsilon_i \geq 0 \\ \forall k, \beta_k \geq 0, \text{and} \sum_{k=1}^{r} \beta_k = 1. \end{cases} \tag{3.20}$$

Here, $\varepsilon_i \geq 0$ are the slack variables, showing the scale of how wrongly sample i is misclassified. ρ represents margin, quantifying the distinction between two classes. The constant factor $D > 0$ is the tradeoff parameter between misclassification error and margin maximization.

In LP-B, the ensemble rules are:

$$f_{\text{ensemble}}(\{\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(r)}\}) = \arg \max_{g\in\{1,\ldots,c\}} \sum_{k=1}^{r} \mathbf{B}_{k,g} \mathbf{h}_k(\mathbf{Y}^{(k)}), \tag{3.21}$$

where $\mathbf{B}_{k,g}$ represents the ensemble weight of the k'th basic classifier for the g'th group. $\mathbf{B}$ is learn by multiclass LPBoost via solving the following optimization:

$$\max_{\boldsymbol{\beta},\rho,\boldsymbol{\xi}} \quad \mathcal{P}(\boldsymbol{\beta},\rho,\boldsymbol{\xi}) = \rho - \frac{1}{\nu n}\sum_{i=1}^{n}\varepsilon_i \tag{3.22}$$

$$\text{subject to} \quad \begin{cases} \forall i, \text{ and } o_j \neq o_i, \ \sum_{k=1}^{r} B_{ko_i}H_{iko_i} - \sum_{k=1}^{r} B_{ko_j}H_{iko_j} \geq \rho - \varepsilon_i, \\ \forall i, \ \varepsilon_i \geq 0 \\ \forall k, g \ B_{kg} \geq 0, \text{and} \sum_{k=1}^{r} B_{kg} = 1, \end{cases} \tag{3.23}$$

The training procedure for LP-β and LP-B are similar. The ideal case is that there is enough data to estimate H_{ikc} and $\boldsymbol{\beta}$ on independent sets. However, in most situation, we need to share data when achieving solutions for basic svm and the ensemble LPBoost. The following two stages scheme are used to avoid biased estimates. First, five fold cross validation (CV) can be performed to select the best hyperparameter for each basic SVM model individually. In this case, this is to find the balancing parameter C_k for each basic multiclass SVM model. At this point the only parameter left is ν. Since there is no independent training data left to set this parameter, we compute for each H_{ikc} the CV output using its best hyperparameter identified before. This results in a prediction for each training point using a classifier which was not trained using that point, but on other 80% of the training data. The CV outputs of all SVMs are used as training data for LP-β. We perform CV to select the best parameter ν and subsequently train the final combination $\boldsymbol{\beta}$. The main concern using this scheme is that the input to the ensemble method is from CV and not from the classifier h_k later used in the combination. However, it is reasonable to assume that the learners used to produce the training data for LP-β are not too different. Comparing LP-β and LP-B, they are both linear programming problem, while LP-B have more parameters so that more expensive to solve than LP-β. Fitting $c \times r$ instead of r parameters demands for more training data, if not, LP-B may results worse model compared to LP-β.

3.3 Simulation Study

To show the performance of our model, simulation study is conducted to exam the performance of SVM and LPBoost. Simulated data are designed in this way: assume there are three disease categories, i.e. $g \in \{0, 1, 2\}$, each of which has 500 samples. Each sample contain $r = 100$ image regions and each region is composed of 10×10 voxels. Denote $\mu_{g,k}, \sigma_{g,k}$ the mean value and standard deviation for group g and region k respectively. $\mathbf{Y}^{(k)}_{10\times10}$ is the random variable representing voxel level signals in region k. Detailed data simulation description is shown in Table (3.1). For the first 25 regions, samples from different disease groups show no difference, so that they have the same regional mean, $\mu_{0,k} = \mu_{1,k} = \mu_{2,k}, k = 1, \ldots, 25$ and the same regional standard deviation, $\sigma_{0,k} = \sigma_{1,k} = \sigma_{2,k}, k = 1, \ldots, 25$. The region mean is randomly sampled from standard Gaussian distribution. The region standard deviation is set to equal to 1 constant. For 26 to 50 regions, image samples from different disease group show difference in the regional mean value, $\mu_{0,k} < \mu_{1,k} < \mu_{2,k}$, but with the same regional standard deviation $\sigma_{0,k} < \sigma_{1,k} = \sigma_{2,k} = 1$. For regions 51 to 75, image samples from different disease group show difference in the regional mean value, $\mu_{0,k} < \mu_{1,k} < \mu_{2,k}$, but with the same standard deviation $\sigma_{0,k} < \sigma_{1,k} = \sigma_{2,k} = 2$. The regional standard deviation in region 51 to 75 is higher than region 26 to 50. For the 76-100 regions, images from different disease group show the same regional mean value, $\mu_{1,k} = \mu_{2,k}, k = 76, \ldots, 100$ but with different variance, $\sigma_{0,k} < \sigma_{1,k} < \sigma_{2,k}, k = 76, \ldots, 100$. The regional standard deviation is randomly drawn from Gaussian distribution with given mean and variance shown in Table(3.1).

Three scenarios are used to simulate three different types of spatial correlation of the images. For every region with a given mean and standard deviation, the voxel signals within a region are generated either random, or following Gaussian Process

Table 3.1: Simulation Setting of Mean and Standard Deviation Values of 100 Regions

Regions	Mean and Standard Deviation	Description
Region 1 ∼ Region 25	$\mu_{0,k} = \mu_{1,k} = \mu_{2,k} \sim \mathcal{N}(0,1)$ $\sigma_{0,k} = \sigma_{1,k} = \sigma_{2,k} = 1$	no difference across three groups
Region 26 ∼ Region 50	$\mu_{0,k} = 0$ $\mu_{1,k} \sim \mathcal{N}(\mu_{0,k} + 0.1, 0.02^2)$ $\mu_{2,k} \sim \mathcal{N}(\mu_{1,k} + 0.1, 0.02^2)$ $\sigma_{0,k} = \sigma_{1,k} = \sigma_{2,k} = 1$	group difference in mean value, same within-region standard deviation
Region 51 ∼ Region 75	$\mu_{0,k} = 0$ $\mu_{1,k} \sim \mathcal{N}(\mu_{0,k} + 0.1, 0.02^2)$ $\mu_{2,k} \sim \mathcal{N}(\mu_{1,k} + 0.1, 0.02^2)$ $\sigma_{0,k} = \sigma_{1,k} = \sigma_{2,k} = 2$	group difference in mean value, same but greater within-region standard deviation
Region 76 ∼ Region 100	$\mu_{0,k} = \mu_{1,k} = \mu_{2,k} = 0$ $\sigma_{0,k} = 1$ $\sigma_{1,k} \sim \mathcal{N}(\sigma_{0,k} + 0.05, 0.01^2)$ $\sigma_{2,k} \sim \mathcal{N}(\sigma_{1,k} + 0.05, 0.01^2)$	group difference in standard deviation, same mean

with low correlation, or following Gaussian Process with moderate correlation. The detailed simulation setting for regional spatial correlation can be found in Table(3.2). Figures (3.1), (3.2) and (3.3) show sample simulated images from different groups ($g = 0, 1, 2$) following these three scenarios.

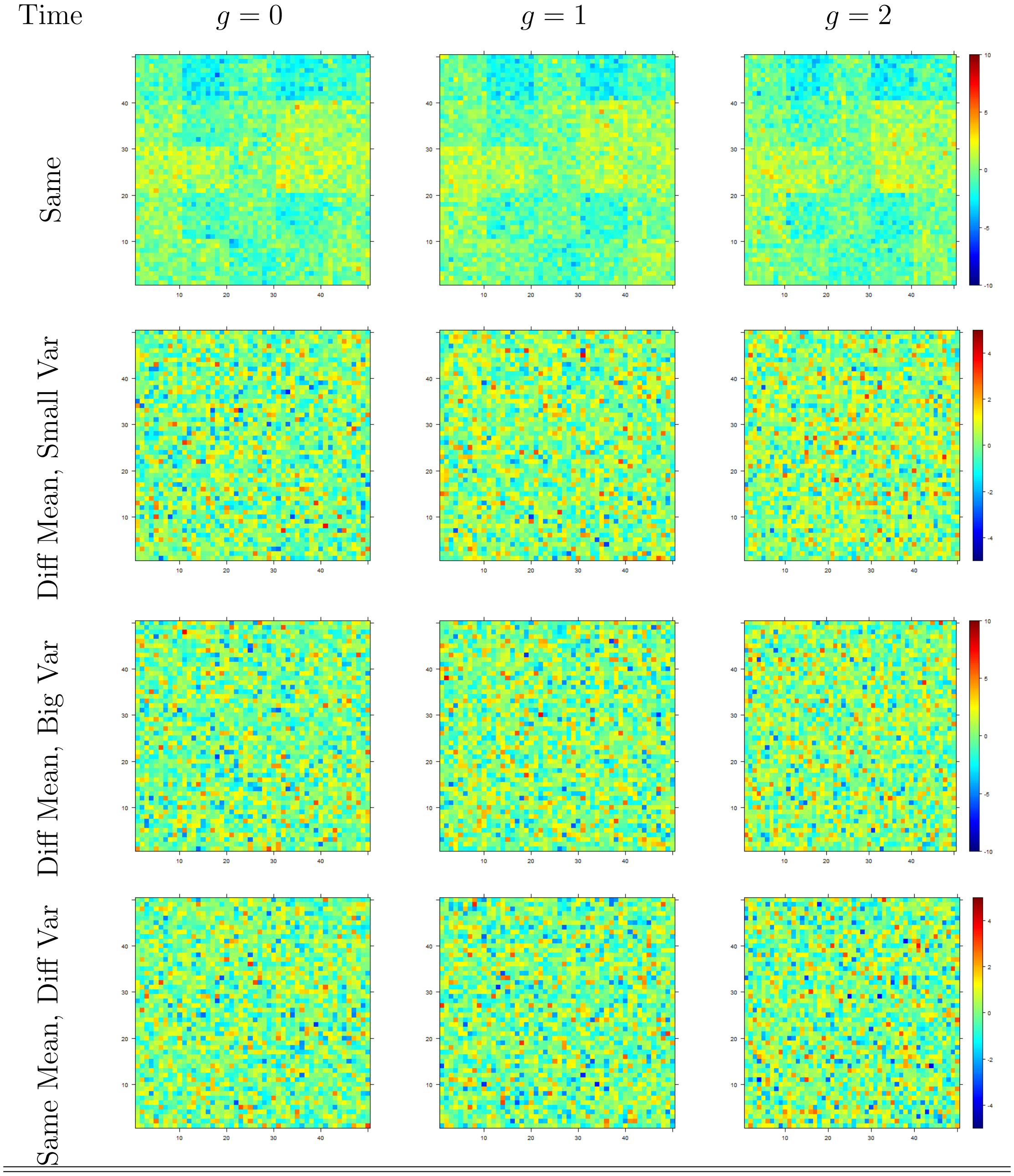
Table 3.2: Three Scenarios of Simulation Study

Scenario I	$\mathbf{Y}^k \sim_{i.i.d} \mathcal{N}(\mu_k, \sigma_k^2)$
Scenario II	$\mathbf{Y}^k \sim \mathcal{GP}(\mu_k, K_{II})$, where $K_{II}(\mathbf{x}, \mathbf{x}') = \sigma_k^2 \exp(-\frac{\lvert\lvert\mathbf{x}-\mathbf{x}'\rvert\rvert^2}{2l_{II}^2})$, $l_{II} = 3$
Scenario III	$\mathbf{Y}^k \sim \mathcal{GP}(\mu_k, K_{III})$, where $K_{III}(\mathbf{x}, \mathbf{x}') = \sigma_k^2 \exp(-\frac{\lvert\lvert\mathbf{x}-\mathbf{x}'\rvert\rvert^2}{2l_{III}^2})$, $l_{III} = 7$

Note: $\mathbf{Y}^k$ denote neuroimaging data of dimension 10 × 10 from region k

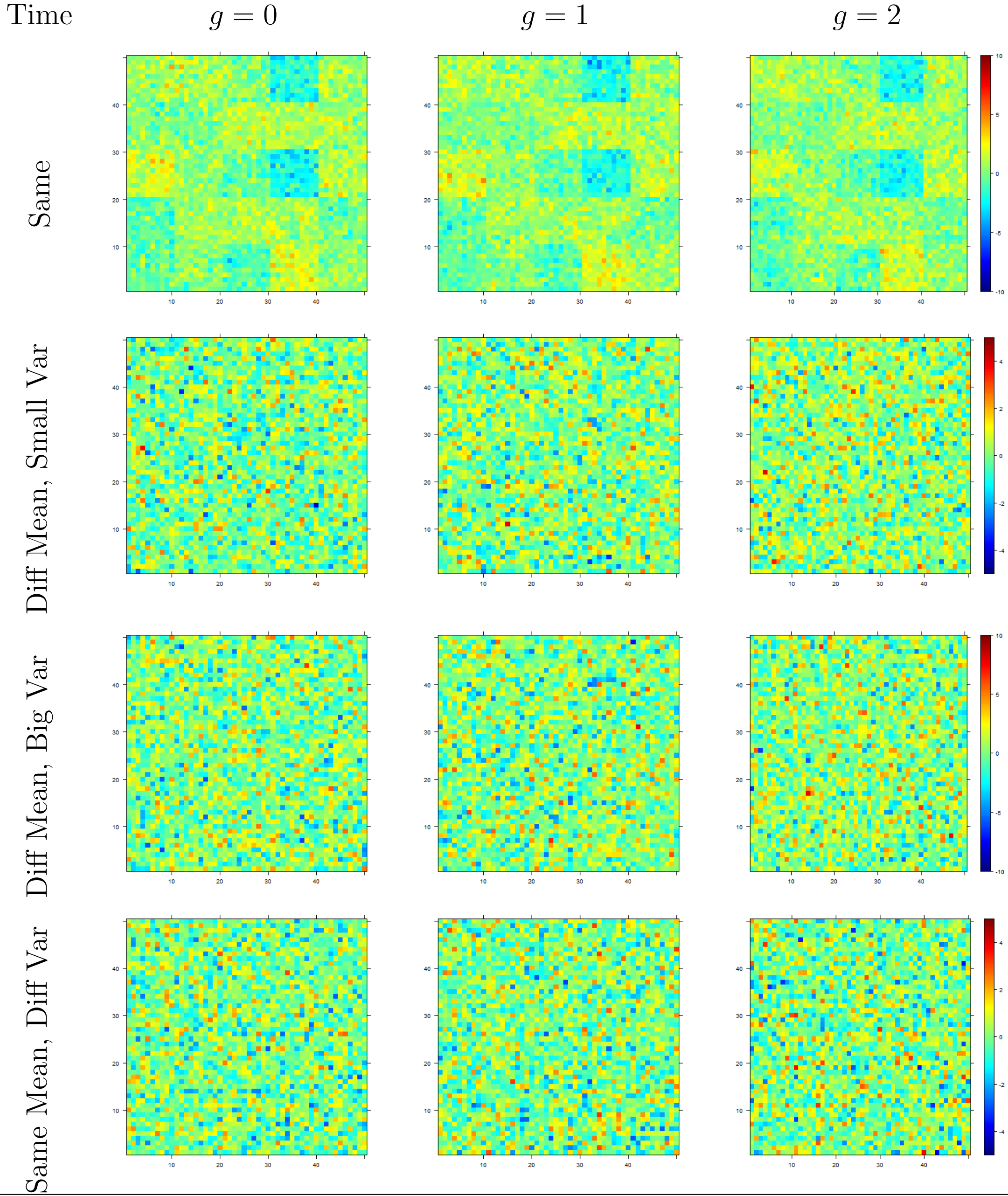
Among all simulated image data, we randomly sample 60% to be training data set, 20% to be valid data set, and the rest 20% to be test data set. Parameters of basic classifiers are learned through training data. Five fold cross validation (CV)

Figure 3.1: One replicate of simulated imaging data of three disease groups, containing 100 regions and each has a 10 × 10 voxel space, the simulated images are generated under senario I, so that the data has no spatial correlation



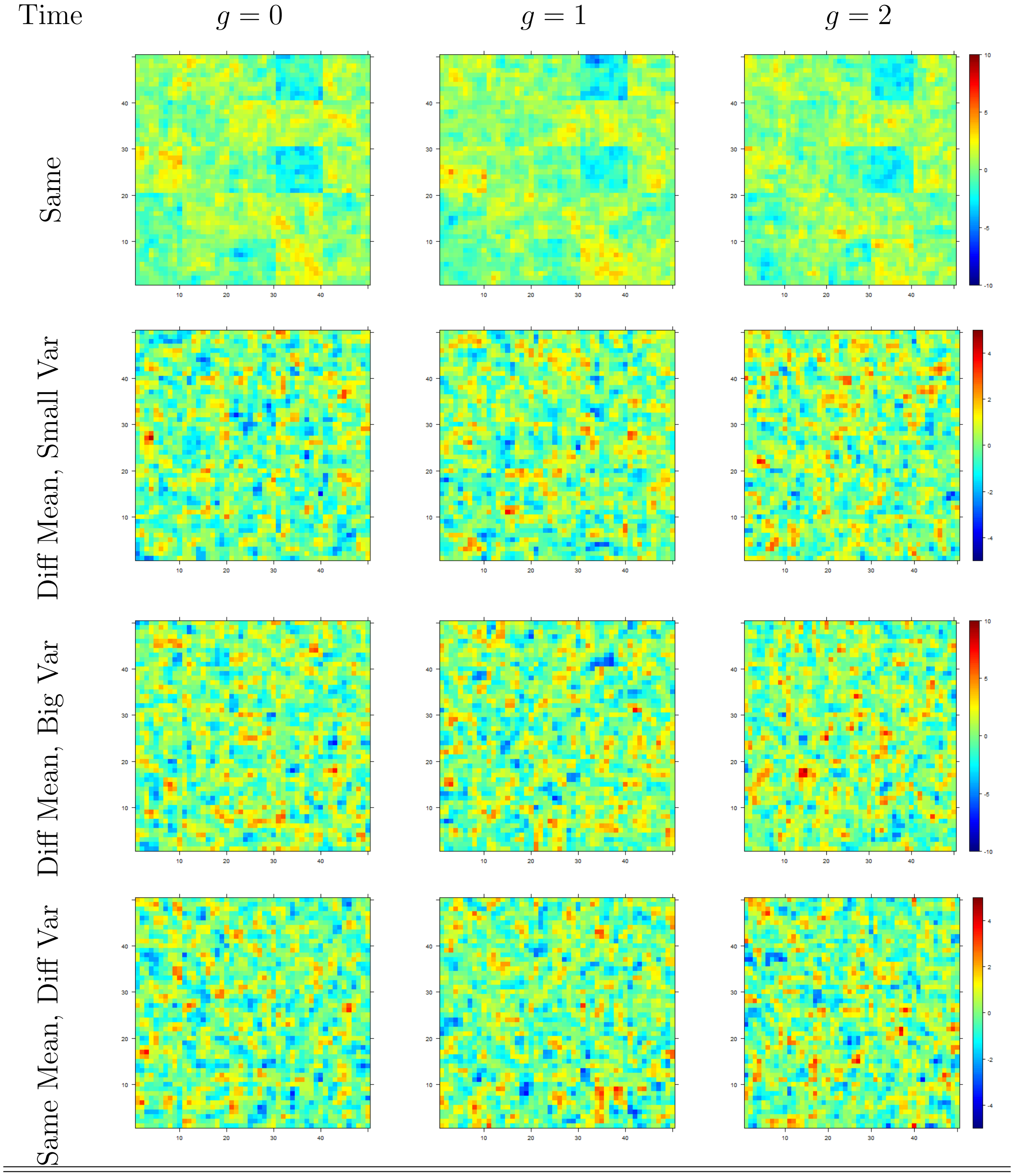
is used in training data set to determine the best tuning parameter C_k for each SVM classifier of region k. The parameters of LPBoost are learned through valid data, using predicted label from trained SVM classifiers as input. Five fold CV is

Figure 3.2: One replicate of simulated imaging data of three disease groups, containing 100 regions and each has a 10×10 voxel space, the simulation images are generated under senario II



used in valid data to determine the best tuning parameter D for LPBoost. The proposed LPBoost method is compared to a baseline SVM classifier (PCA+SVM), which is built based on principle components extracted over all regions combining

Figure 3.3: One replicate of simulated imaging data of three disease groups, containing 100 regions and each has a 10×10 voxel space, the simulation images are generated under senario III



train and valid data sets together. All predictive errors are calculated through test data. Table(3.3) and Table(3.4) show the simulation results of binary classification and three group classification respectively. The results are based on one random split of train, valid and test data set. Table(3.5) and Table (3.6) show simulation results of LPBoost weights of different regions under three simulation scenarios based on the same train, valid, data set separation. Ranking based on LPBoost weight and also the SVM predictive errors are also provided.

Table 3.3: Result summary of simulation study for binary classification. The test errors of basic classifier SVM, PCA+SVM, and LPBoost methods are listed based on one random Train/Valid/Test split

	Regional SVM, mean (sd) [min, max]				PCA+SVM	LPBoost
	Region 1∼25	Region 26∼50	Region 51∼75	Region 76∼100	All Region	All Region
Scenario I	0.501(0.028) [0.430, 0.555]	0.358(0.055) [0.260, 0.450]	0.463(0.035) [0.385, 0.515]	0.417(0.049) [0.315, 0.510]	0.218	0.135
Scenario II	0.500(0.026) [0.445, 0.565]	0.376(0.045) [0295, 0.460]	0.454(0.034) [0.385, 0.515]	0.411(0.033) [0.355, 0.470]	0.201	0.140
Scenario III	0.494(0.024) [0.435, 0.530]	0.435(0.037) [0.370, 0.510]	0.484(0.028) [0.420, 0.525]	0.458(0.034) [0.400,0.540]	0.357	0.300

Table 3.4: Result summary of simulation study for three group classification. The test errors of basic classifier SVM, PCA+SVM, and LPBoost methods are listed based on one random Train/Valid/Test split

	Regional SVM, mean (sd) [min, max]				PCA+SVM	LPBoost
	Region 1∼25	Region 26∼50	Region 51∼75	Region 76∼100	All Region	All Region
Scenario I	0.676(0.030) [0.620, 0.735]	0.478(0.051) [0.39, 0.580]	0.586(0.049) [0.475, 0.665]	0.556(0.038) [0.505, 0.635]	0.311	0.253
Scenario II	0.668(0.037) [0.595, 0.735]	0.474(0.036) [0.395, 0.530]	0.587(0.043) [0.485, 0.690]	0.550(0.039) [0.480, 0.665]	0.354	0.289
Scenario III	0.672(0.035) [0.600, 0.725]	0.578(0.040) [0.460, 0.640]	0.644(0.026) [0.580, 0.685]	0.591(0.033) [0.515,0.655]	0.396	0.344

In order to further test the performance of proposed LPBoost method, the experiments are conducted 50 times based on random splits of train, valid, and test sets. Table (3.7) shows the mean and standard error of the classification error of LPBoost, the classification error of PCA+SVM, the ratio of the classification error of LPBoost divided by the best regional SVM classifier, and the ratio of the classification error of LPBoost divided by the one of PCA+SVM. As a result, the proposed LPBoost method has smaller classification error under all three scenarios, along with smaller standard error. For binary classification, the proposed LPBoost method decreases the classification error by around 43% through ensemble for scenario I and II, and around

Table 3.5: Result summary of simulation study for binary classification. The relative importance of different regions (one random Train/Valid/Test split) based on basic SVM classifiers and on LPBoost are listed.

	Scenario I				Scenario II				Scenario III			
Region	LPBoost		SVM		LPBoost		SVM		LPBoost		SVM	
#	order	coef	order	pe	order	coef	order	pe	order	coef	order	pe
1	66	0.000	74	0.495	72	0.000	86	0.500	70	0.000	76	0.500
2	66	0.000	60	0.470	72	0.000	80	0.490	70	0.000	52	0.475
3	66	0.000	83	0.500	32	0.011	94	0.505	70	0.000	96	0.520
4	66	0.000	98	0.540	72	0.000	96	0.520	70	0.000	76	0.500
5	66	0.000	96	0.530	72	0.000	86	0.500	70	0.000	56	0.480
6	66	0.000	98	0.540	72	0.000	86	0.500	70	0.000	76	0.500
7	66	0.000	83	0.500	72	0.000	100	0.565	70	0.000	93	0.515
8	66	0.000	83	0.500	72	0.000	86	0.500	70	0.000	76	0.500
9	66	0.000	100	0.555	72	0.000	55	0.440	70	0.000	76	0.500
10	66	0.000	83	0.500	72	0.000	86	0.500	70	0.000	76	0.500
11	66	0.000	83	0.500	72	0.000	74	0.480	70	0.000	41	0.460
12	66	0.000	83	0.500	72	0.000	86	0.500	70	0.000	76	0.500
13	66	0.000	83	0.500	72	0.000	60	0.455	31	0.008	46	0.465
14	66	0.000	95	0.520	72	0.000	98	0.525	33	0.007	64	0.495
15	66	0.000	83	0.500	72	0.000	96	0.520	70	0.000	76	0.500
16	66	0.000	92	0.510	72	0.000	86	0.500	38	0.002	93	0.515
17	66	0.000	98	0.540	72	0.000	64	0.460	70	0.000	76	0.500
18	66	0.000	83	0.500	72	0.000	64	0.460	70	0.000	76	0.500
19	66	0.000	54	0.450	72	0.000	86	0.500	70	0.000	23	0.435
20	66	0.000	83	0.500	72	0.000	60	0.455	70	0.000	98	0.530
21	66	0.000	83	0.500	72	0.000	74	0.480	70	0.000	50	0.470
22	66	0.000	44	0.430	72	0.000	86	0.500	70	0.000	98	0.530
23	66	0.000	83	0.500	72	0.000	86	0.500	70	0.000	32	0.445
24	66	0.000	65	0.475	72	0.000	86	0.500	70	0.000	76	0.500
25	66	0.000	58	0.465	72	0.000	99	0.540	70	0.000	96	0.520
26	15	0.030	6	0.305	72	0.000	10	0.345	22	0.012	36	0.450
27	21	0.017	23	0.380	72	0.000	16	0.370	18	0.017	15	0.425
28	23	0.014	23	0.380	20	0.023	26	0.385	25	0.009	46	0.465
29	66	0.000	12	0.340	23	0.019	21	0.380	70	0.000	46	0.465
30	20	0.022	18	0.370	10	0.034	21	0.380	27	0.009	46	0.465
31	66	0.000	46	0.435	33	0.011	7	0.335	5	0.063	10	0.415
32	66	0.000	12	0.340	18	0.028	60	0.455	70	0.000	36	0.450
33	2	0.079	10	0.330	13	0.032	21	0.380	70	0.000	36	0.450
34	6	0.060	3	0.285	35	0.010	34	0.405	6	0.056	88	0.505
35	18	0.026	32	0.405	5	0.046	14	0.365	70	0.000	15	0.425
36	12	0.035	30	0.400	40	0.003	36	0.410	70	0.000	23	0.435
37	3	0.078	14	0.360	19	0.023	30	0.395	24	0.011	15	0.425
38	66	0.000	54	0.450	25	0.017	26	0.385	32	0.007	28	0.440
39	8	0.047	11	0.335	72	-0.000	1	0.280	36	0.006	1	0.370
40	66	0.000	36	0.415	1	0.077	8	0.340	16	0.025	23	0.435
41	1	0.084	5	0.295	72	0.000	3	0.305	70	0.000	5	0.395
42	28	0.003	20	0.375	72	0.000	4	0.320	70	0.000	4	0.390
43	30	0.000	36	0.415	4	0.047	6	0.330	13	0.029	8	0.405
44	7	0.054	1	0.260	8	0.035	2	0.300	70	0.000	2	0.385
45	5	0.068	54	0.450	72	0.000	56	0.445	70	0.000	36	0.450
46	17	0.029	8	0.315	9	0.034	40	0.415	70	0.000	91	0.510
47	14	0.032	2	0.270	72	0.000	10	0.345	14	0.029	18	0.430
48	66	0.000	16	0.365	3	0.047	6	0.330	12	0.033	2	0.385
49	31	0.000	20	0.375	72	-0.000	46	0.425	7	0.049	76	0.500
50	11	0.040	4	0.290	6	0.043	14	0.365	20	0.016	10	0.415

80% through ensemble for scenario III. Comparing to the PCA+SVM method conducting on all regions, the proposed LPBoost shows better performance which has an about 60% smaller classification error for scenario I and II, and about 85% smaller for

Table 3.6: Continued to 3.5

	Scenario I				Scenario II				Scenario III			
Region	LPBoost		SVM		LPBoost		SVM		LPBoost		SVM	
#	order	coef	order	pe	order	coef	order	pe	order	coef	order	pe
51	66	0.000	74	0.495	72	0.000	58	0.450	70	0.000	64	0.495
52	66	0.000	83	0.500	16	0.029	58	0.450	70	0.000	59	0.485
53	66	0.000	44	0.430	72	0.000	78	0.485	70	0.000	76	0.500
54	66	0.000	40	0.425	72	0.000	94	0.505	70	0.000	76	0.500
55	66	0.000	46	0.435	72	0.000	40	0.415	70	0.000	18	0.430
56	66	0.000	65	0.475	72	0.000	34	0.405	39	0.000	18	0.430
57	66	0.000	49	0.445	72	0.000	68	0.465	70	0.000	93	0.515
58	66	0.000	57	0.460	39	0.003	40	0.415	70	0.000	56	0.480
59	66	0.000	65	0.475	72	0.000	74	0.480	28	0.009	62	0.490
60	66	0.000	72	0.485	72	0.000	70	0.470	70	0.000	76	0.500
61	66	0.000	83	0.500	72	0.000	60	0.455	70	0.000	28	0.440
62	66	0.000	26	0.385	30	0.013	46	0.425	70	0.000	64	0.495
63	66	0.000	83	0.500	11	0.033	64	0.460	10	0.039	59	0.485
64	66	0.000	60	0.470	36	0.007	52	0.435	21	0.015	32	0.445
65	66	0.000	65	0.475	43	0.002	52	0.435	70	0.000	76	0.500
66	66	0.000	65	0.475	28	0.015	78	0.485	70	0.000	76	0.500
67	66	0.000	94	0.515	41	0.002	68	0.465	70	0.000	52	0.475
68	66	0.000	83	0.500	29	0.014	95	0.510	70	0.000	97	0.525
69	66	0.000	29	0.395	72	0.000	46	0.425	70	0.000	64	0.495
70	66	0.000	72	0.485	14	0.030	26	0.385	29	0.008	13	0.420
71	66	0.000	39	0.420	72	0.000	86	0.500	70	0.000	76	0.500
72	29	0.001	44	0.430	17	0.029	36	0.410	70	0.000	76	0.500
73	66	0.000	54	0.450	72	0.000	74	0.480	23	0.011	56	0.480
74	66	0.000	83	0.500	7	0.039	71	0.475	70	0.000	88	0.505
75	66	0.000	49	0.445	15	0.030	49	0.430	70	0.000	88	0.505
76	66	0.000	26	0.385	72	0.000	49	0.430	15	0.028	62	0.490
77	26	0.004	28	0.390	31	0.013	18	0.375	70	0.000	23	0.435
78	66	0.000	20	0.375	72	0.000	52	0.435	70	0.000	18	0.430
79	66	0.000	8	0.315	72	0.000	40	0.415	70	0.000	52	0.475
80	66	0.000	16	0.365	42	0.002	32	0.400	8	0.046	32	0.445
81	66	0.000	28	0.390	12	0.032	28	0.390	3	0.071	36	0.450
82	66	0.000	36	0.415	2	0.054	30	0.395	1	0.092	46	0.465
83	10	0.045	14	0.360	37	0.005	32	0.400	37	0.005	10	0.415
84	9	0.045	40	0.425	72	0.000	44	0.420	70	0.000	46	0.465
85	24	0.007	65	0.475	21	0.021	49	0.430	70	0.000	76	0.500
86	66	0.000	44	0.430	34	0.011	10	0.345	2	0.071	28	0.440
87	66	0.000	34	0.410	22	0.020	21	0.380	4	0.064	23	0.435
88	25	0.005	70	0.480	38	0.005	44	0.420	70	0.000	10	0.415
89	66	0.000	49	0.445	72	0.000	21	0.380	30	0.008	36	0.450
90	4	0.068	30	0.400	72	0.000	68	0.465	70	0.000	41	0.460
91	13	0.033	10	0.330	24	0.019	12	0.355	26	0.009	6	0.400
92	66	0.000	23	0.380	72	0.000	74	0.480	70	0.000	59	0.485
93	22	0.016	34	0.410	26	0.017	26	0.385	9	0.045	28	0.440
94	66	0.000	58	0.465	27	0.015	40	0.415	35	0.006	56	0.480
95	16	0.030	65	0.475	72	0.000	14	0.365	11	0.037	8	0.405
96	27	0.003	36	0.415	72	0.000	40	0.415	19	0.016	32	0.445
97	66	0.000	92	0.510	72	0.000	68	0.465	70	0.000	88	0.505
98	66	0.000	54	0.450	72	0.000	18	0.375	17	0.024	41	0.460
99	66	0.000	70	0.480	72	0.000	74	0.480	34	0.007	100	0.540
100	19	0.024	54	0.450	72	0.000	52	0.435	70	0.000	50	0.470

scenario III. For three group classification, the proposed LPBoost methods decreases the classification error by 64.9%, 73.2%, and 74.8% through ensemble for scenario I, II, and III respectively. Comparing to the PCA+SVM method conducting on all regions, the proposed LPBoost shows better performance which has an about 82.5%, 83.5%, and 85.2% smaller classification error for scenario I, II, and III respectively.

Table 3.7: Simulation results for binary classification and three group classification comparing LPBoost method with PCA+SVM method and regional SVM methods based on 50 repeated experiments, randomly spliting train, valid, test data sets. The measurement metrics includ classification error and its standard error, classification error ratio between two methods of interest

		Scenario I	Scenario II	Scenario III
Binary Classification	LPBoost	0.123(0.034)	0.119(0.034)	0.292(0.040)
	PCA+SVM	0.218(0.042)	0.192(0.037)	0.354(0.041)
	LPBoost/PCA+SVM	0.434(0.105)	0.424(0.111)	0.808(0.101)
	LPBoost/best regional SVM	0.594(0.213)	0.637(0.206)	0.846(0.152)
Three Group Classification	LPBoost	0.255(0.045)	0.289(0.051)	0.344(0.053)
	PCA+SVM	0.309(0.048)	0.346(0.050)	0.403(0.054)
	LPBoost/PCA+SVM	0.649(0.124)	0.732(0.118)	0.7486(0.130)
	LPBoost/best regional SVM	0.825(0.210)	0.835(0.232)	0.853(0.226)

3.4 ABIDE Data Analysis

Autism spectrum disorder (ASD) is a widely recognized disease characterized by qualitative impairment in social reciprocity, and by repetitive, restricted, and stereotyped behaviors. Due to its high prevalence in children with a more than 1% occurrence rate, there is strong need to further understand the mechanisms underlying ASD in order to identify ways of earlier diagnosis, optional treatment selection, and better outcome prediction. Autism Brain Imaging Data Exchange (ABIDE) is a consortium of the International Neuroimaging Datasharing Initiative. ABIDE collaborated 16 international imaging sites and collected neuroimaging data from 539 individuals suffering from ASD and 573 typical controls (TC). The datasets are composed of structural and resting state functional MRI data along with an extensive array of prototypical information. The major goal of ABIDE is to provide data support to accelerate research of the neural based of ASD (*Di Martino et al.*, 2014).

In our preprocessed data, we refer to *Woods et al.* (1998) for the details on the preprocessing steps for MRI data. In the end, we keep 514 individuals with ASD and 557 TCs by removing the subjects with low quality on imaging data or having a large proportion of the missing values. We used only resting state fMRI data and built classification model to detect patients' ASD status (ASD: 1, and TC: 0). The data contains fractional Amplitude of Low Frequency Fluctuations (fALFF) measurements on a total of 175,493 voxels (*Zou et al.*, 2008). The fALFF values ranging from 0 to 1. There a total of 116 different regions. The region partition system is based on AAL (the Anatomical Automatic Labeling System) (*Tzourio-Mazoyer et al.*, 2002). Binary classification are performed using SVM. Individual basic classifiers are trained on 116 regions.

There are three experiments conducted based on different train/valid/test set splitting rules. In experiment I, stratified random sample is used to draw train, valid, and test data sets per experimental sites following percentages of 60%, 20%, and 20% respectively. In experiment II, all experiment sites are ranked by their number of subjects in decreasing order and then the first 8 hospitals are selected for training (about 63% of overall subjects), the next 6 hospitals are selected for valid (about 21%), and the last 6 hospitals are selected for test (about 16%). In experiment III, all experiment sites are ranked by their number of subjects in increasing order, as a result, 16 hospitals are selected for training (about 57% of overall subjects), 3 for valid (about 26%), and 1 for testing (17%). Details about how the three experiments are held can be found in Table (3.8, 3.9, and 3.10). The proposed LPBoost method is applied to ABIDE study under three experiments. We achieved a classification accuracy of about 82.4% to classify ASD and TC in experiment I, 84.4% in experiment II (Table 3.11), which is higher than the current best results of 67% by *Abraham et al.* (2017) and 70% by **?** using the same ABIDE study data. The relative importance of different regions for disease diagnostic can be reflected by the associated ensemble weight parameters. The relative importance analysis of different brain regions can be found in the Table (3.12), (3.14), and (3.15).

Table 3.8: A summary of stratified random sampling conducted for ABIDE data in experiment I

	SITE_ID	#	# Train	% Train	# Valid	%Valid	# Test	%Test
1	CALTECH	38	23	0.605	8	0.211	7	0.184
2	CMU	26	16	0.615	6	0.231	4	0.154
3	KKI	45	27	0.600	9	0.200	9	0.200
4	LEUVEN_1	28	17	0.607	6	0.214	5	0.179
5	LEUVEN_2	34	21	0.618	7	0.206	6	0.176
6	MAX_MUN	56	34	0.607	12	0.214	10	0.179
7	NYU	182	110	0.604	37	0.203	35	0.192
8	OHSU	25	15	0.600	5	0.200	5	0.200
9	OLIN	35	21	0.600	7	0.200	7	0.200
10	PITT	56	34	0.607	12	0.214	10	0.179
11	SBL	30	18	0.600	6	0.200	6	0.200
12	SDSU	35	21	0.600	7	0.200	7	0.200
13	STANFORD	40	24	0.600	8	0.200	8	0.200
14	TRINITY	48	29	0.604	10	0.208	9	0.188
15	UCLA_1	73	44	0.603	15	0.205	14	0.192
16	UCLA_2	24	15	0.625	5	0.208	4	0.167
17	UM_1	110	66	0.600	22	0.200	22	0.200
18	UM_2	34	21	0.618	7	0.206	6	0.176
19	USM	96	58	0.604	20	0.208	18	0.188
20	YALE	56	34	0.607	12	0.214	10	0.179
	Total	1071	648	0.605	221	0.206	202	0.189

Table 3.9: A summary of Train/Test/Valid split of ABIDE data in experiment II

	SITE_ID	# of Sample	Total Sample	Total Percentage
train	NYU	182	677	0.632
	UM_1	110		
	USM	96		
	UCLA_1	73		
	MAX_MUN	56		
	PITT	56		
	YALE	56		
	TRINITY	48		
valid	KKI	45	227	0.212
	STANFORD	40		
	CALTECH	38		
	OLIN	35		
	SDSU	35		
	LEUVEN_2	34		
test	UM_2	34	167	0.156
	SBL	30		
	LEUVEN_1	28		
	CMU	26		
	OHSU	25		
	UCLA_2	24		

Table 3.10: A summary of Train/Test/Valid split of ABIDE data in experiment III

	SITE_ID	# of Sample	Total Sample	Total Percentage
train	UCLA_2	24	610	0.570
	OHSU	25		
	CMU	26		
	LEUVEN_1	28		
	SBL	30		
	LEUVEN_2	34		
	UM_2	34		
	OLIN	35		
	SDSU	35		
	CALTECH	38		
	STANFORD	40		
	KKI	45		
	TRINITY	48		
	MAX_MUN	56		
	PITT	56		
	YALE	56		
valid	UCLA_1	73	279	0.261
	USM	96		
	UM_1	110		
test	NYU	182	182	0.170

Table 3.11: LPBoost classification error under three experiments. Results for experiment I is based on 50 replicates of stratified random sample of Train/Valid/Test data sets

	Experiment I, mean(sd)	Experiment II	Experiment II
LPBoost Classification Error	0.176(0.057)	0.156	0.379
LPBoost Type I Error	0.148(0.052)	0.130	0.450
LPBoost Type II Error	0.191(0.060)	0.174	0.321

Table 3.12: A result summary of relative importance of different brain regions for ABIDE analysis based on 50 repeated stratified random samples under Experiment I

LPBoost					SVM	
Rank	Region #	Region Code	Region Name	Coefficient	Rank	PE
1	89	8301	Temporal_Inf_L	0.068(0.194)	19	0.383(0.031)
2	90	8302	Temporal_Inf_R	0.047(0.139)	1	0.353(0.033)
3	56	5402	Fusiform_R	0.044(0.139)	7	0.369(0.032)
4	2	2002	Precentral_R	0.040(0.140)	13	0.374(0.037)
5	58	6002	Postcentral_R	0.037(0.140)	8	0.370(0.029)
6	44	5002	Calcarine_R	0.026(0.141)	24	0.390(0.028)
7	85	8201	Temporal_Mid_L	0.026(0.024)	3	0.362(0.031)
8	1	2001	Precentral_L	0.026(0.022)	2	0.358(0.031)
9	30	3002	Insula_R	0.023(0.021)	10	0.371(0.026)
10	86	8202	Temporal_Mid_R	0.020(0.021)	4	0.367(0.031)
11	92	9002	Cerebelum_Crus1_R	0.018(0.019)	12	0.373(0.035)
12	55	5401	Fusiform_L	0.018(0.018)	9	0.370(0.028)
13	68	6302	Precuneus_R	0.017(0.023)	22	0.388(0.028)
14	67	6301	Precuneus_L	0.017(0.018)	5	0.367(0.030)
15	4	2102	Frontal_Sup_R	0.016(0.020)	6	0.369(0.032)
16	103	9061	Cerebelum_8_L	0.015(0.015)	35	0.398(0.033)
17	82	8112	Temporal_Sup_R	0.014(0.016)	27	0.392(0.026)
18	59	6101	Parietal_Sup_L	0.014(0.016)	14	0.377(0.036)
19	91	9001	Cerebelum_Crus1_L	0.013(0.017)	15	0.378(0.034)
20	20	2402	Supp_Motor_Area_R	0.013(0.016)	26	0.392(0.035)
21	57	6001	Postcentral_L	0.012(0.018)	16	0.378(0.031)
22	29	3001	Insula_L	0.012(0.016)	34	0.397(0.038)
23	23	2601	Frontal_Sup_Medial_L	0.012(0.021)	17	0.379(0.036)
24	7	2201	Frontal_Mid_L	0.012(0.015)	20	0.386(0.036)
25	88	8212	Temporal_Pole_Mid_R	0.011(0.016)	67	0.420(0.036)
26	16	2322	Frontal_Inf_Orb_R	0.011(0.015)	21	0.387(0.029)
27	3	2101	Frontal_Sup_L	0.011(0.016)	32	0.395(0.027)
28	28	2702	Rectus_R	0.011(0.014)	68	0.421(0.036)
29	13	2311	Frontal_Inf_Tri_L	0.010(0.014)	39	0.399(0.034)
30	43	5001	Calcarine_L	0.010(0.017)	41	0.400(0.030)
31	94	9012	Cerebelum_Crus2_R	0.010(0.014)	39	0.399(0.037)
32	8	2202	Frontal_Mid_R	0.010(0.013)	12	0.373(0.031)
33	87	8211	Temporal_Pole_Mid_L	0.010(0.015)	39	0.399(0.030)
34	18	2332	Rolandic_Oper_R	0.010(0.014)	54	0.411(0.034)
35	70	6402	Paracentral_Lobule_R	0.009(0.017)	62	0.415(0.028)
36	50	5102	Occipital_Sup_R	0.009(0.013)	23	0.389(0.027)
37	38	4102	Hippocampus_R	0.009(0.014)	78	0.427(0.035)
38	64	6212	SupraMarginal_R	0.009(0.013)	42	0.402(0.033)
39	19	2401	Supp_Motor_Area_L	0.009(0.014)	36	0.398(0.030)
40	78	7102	Thalamus_R	0.009(0.013)	55	0.411(0.031)
41	26	2612	Frontal_Med_Orb_R	0.008(0.012)	58	0.412(0.032)
42	40	4112	ParaHippocampal_R	0.008(0.014)	31	0.394(0.033)
43	93	9011	Cerebelum_Crus2_L	0.008(0.014)	51	0.408(0.031)
44	39	4111	ParaHippocampal_L	0.007(0.012)	47	0.406(0.031)
45	99	9041	Cerebelum_6_L	0.007(0.012)	25	0.390(0.032)
46	34	4012	Cingulum_Mid_R	0.007(0.011)	33	0.396(0.033)
47	15	2321	Frontal_Inf_Orb_L	0.007(0.012)	43	0.403(0.032)
48	83	8121	Temporal_Pole_Sup_L	0.007(0.014)	56	0.412(0.033)
49	81	8111	Temporal_Sup_L	0.007(0.010)	44	0.405(0.034)
50	6	2112	Frontal_Sup_Orb_R	0.007(0.012)	49	0.407(0.034)
51	31	4001	Cingulum_Ant_L	0.007(0.012)	66	0.419(0.029)
52	63	6211	SupraMarginal_L	0.007(0.012)	60	0.414(0.033)
53	32	4002	Cingulum_Ant_R	0.007(0.013)	58	0.412(0.037)
54	97	9031	Cerebelum_4_5_L	0.006(0.012)	72	0.422(0.034)
55	51	5201	Occipital_Mid_L	0.006(0.011)	18	0.382(0.031)
56	5	2111	Frontal_Sup_Orb_L	0.006(0.011)	65	0.418(0.039)
57	48	5022	Lingual_R	0.006(0.010)	28	0.392(0.030)
58	106	9072	Cerebelum_9_R	0.006(0.011)	89	0.441(0.043)
59	27	2701	Rectus_L	0.006(0.012)	69	0.421(0.036)
60	84	8122	Temporal_Pole_Sup_R	0.006(0.012)	63	0.417(0.036)

Table 3.13: Continued to 3.12

LPBoost					SVM	
Rank	Region #	Region Code	Region Name	Coefficient	Rank	PE
61	104	9062	Cerebelum_8_R	0.006(0.011)	53	0.409(0.040)
62	61	6201	Parietal_Inf_L	0.005(0.009)	46	0.405(0.028)
63	47	5021	Lingual_L	0.005(0.009)	52	0.408(0.034)
64	17	2331	Rolandic_Oper_L	0.005(0.012)	62	0.415(0.029)
65	74	7012	Putamen_R	0.005(0.010)	76	0.425(0.037)
66	33	4011	Cingulum_Mid_L	0.005(0.009)	58	0.412(0.032)
67	111	9120	Vermis_4_5	0.005(0.009)	72	0.422(0.033)
68	37	4101	Hippocampus_L	0.005(0.012)	48	0.407(0.038)
69	24	2602	Frontal_Sup_Medial_R	0.005(0.009)	30	0.393(0.033)
70	9	2211	Frontal_Mid_Orb_L	0.005(0.012)	74	0.424(0.034)
71	66	6222	Angular_R	0.005(0.009)	75	0.424(0.035)
72	69	6401	Paracentral_Lobule_L	0.005(0.011)	73	0.424(0.035)
73	98	9032	Cerebelum_4_5_R	0.004(0.009)	70	0.422(0.034)
74	100	9042	Cerebelum_6_R	0.004(0.008)	37	0.399(0.032)
75	105	9071	Cerebelum_9_L	0.004(0.009)	95	0.452(0.034)
76	49	5101	Occipital_Sup_L	0.004(0.009)	77	0.425(0.032)
77	10	2212	Frontal_Mid_Orb_R	0.004(0.010)	79	0.428(0.034)
78	52	5202	Occipital_Mid_R	0.004(0.010)	45	0.405(0.035)
79	102	9052	Cerebelum_7b_R	0.004(0.008)	91	0.444(0.034)
80	46	5012	Cuneus_R	0.004(0.010)	64	0.418(0.034)
81	72	7002	Caudate_R	0.004(0.008)	93	0.447(0.031)
82	71	7001	Caudate_L	0.003(0.009)	86	0.437(0.038)
83	12	2302	Frontal_Inf_Oper_R	0.003(0.008)	50	0.408(0.032)
84	65	6221	Angular_L	0.003(0.008)	81	0.429(0.036)
85	77	7101	Thalamus_L	0.003(0.008)	85	0.434(0.030)
86	14	2312	Frontal_Inf_Tri_R	0.003(0.007)	80	0.429(0.036)
87	60	6102	Parietal_Sup_R	0.003(0.007)	29	0.393(0.037)
88	54	5302	Occipital_Inf_R	0.003(0.008)	88	0.439(0.028)
89	53	5301	Occipital_Inf_L	0.003(0.008)	90	0.441(0.036)
90	25	2611	Frontal_Med_Orb_L	0.003(0.008)	84	0.433(0.034)
91	22	2502	Olfactory_R	0.002(0.006)	96	0.456(0.039)
92	62	6202	Parietal_Inf_R	0.002(0.007)	82	0.432(0.025)
93	73	7011	Putamen_L	0.002(0.007)	92	0.446(0.030)
94	11	2301	Frontal_Inf_Oper_L	0.002(0.006)	97	0.462(0.034)
95	101	9051	Cerebelum_7b_L	0.001(0.005)	87	0.437(0.032)
96	112	9130	Vermis_6	0.001(0.004)	94	0.450(0.035)
97	42	4202	Amygdala_R	0.001(0.005)	106	0.485(0.031)
98	75	7021	Pallidum_L	0.001(0.004)	105	0.485(0.032)
99	108	9082	Cerebelum_10_R	0.001(0.005)	101	0.476(0.029)
100	113	9140	Vermis_7	0.001(0.005)	113	0.495(0.036)
101	110	9110	Vermis_3	0.001(0.003)	99	0.471(0.037)
102	95	9021	Cerebelum_3_L	0.001(0.004)	116	0.505(0.031)
103	45	5011	Cuneus_L	0.001(0.005)	83	0.432(0.033)
104	116	9170	Vermis_10	0.001(0.005)	115	0.501(0.033)
105	96	9022	Cerebelum_3_R	0.001(0.003)	107	0.488(0.031)
106	41	4201	Amygdala_L	0.001(0.004)	108	0.490(0.035)
107	36	4022	Cingulum_Post_R	0.000(0.002)	111	0.493(0.031)
108	79	8101	Heschl_L	0.000(0.002)	102	0.477(0.040)
109	76	7022	Pallidum_R	0.000(0.002)	104	0.482(0.035)
110	35	4021	Cingulum_Post_L	0.000(0.002)	100	0.472(0.036)
111	21	2501	Olfactory_L	0.000(0.001)	98	0.470(0.035)
112	109	9100	Vermis_1_2	0.000(0.001)	103	0.481(0.031)
113	80	8102	Heschl_R	0.000(0.001)	112	0.493(0.036)
114	107	9081	Cerebelum_10_L	0.000(0.001)	109	0.491(0.033)
115	114	9150	Vermis_8	0.000(0.000)	114	0.498(0.036)
116	115	9160	Vermis_9	0.000(0.000)	110	0.491(0.033)

Table 3.14: A result summary of relative importance of different brain regions for ABIDE analysis based on Experiment II

LPBoost					SVM	
Rank	Region #	Region Code	Region Name	Coefficient	Rank	PE
1	37	4101	Hippocampus_L	0.054	64.0	0.425
2	64	6212	SupraMarginal_R	0.043	40.5	0.401
3	56	5402	Fusiform_R	0.042	71.5	0.437
4	60	6102	Parietal_Sup_R	0.040	47.0	0.407
5	2	2002	Precentral_R	0.039	40.5	0.401
6	43	5001	Calcarine_L	0.035	22.0	0.377
7	65	6221	Angular_L	0.030	71.5	0.437
8	90	8302	Temporal_Inf_R	0.030	22.0	0.377
9	72	7002	Caudate_R	0.026	93.0	0.461
10	23	2601	Frontal_Sup_Medial_L	0.025	3.5	0.341
11	48	5022	Lingual_R	0.025	53.5	0.413
12	1	2001	Precentral_L	0.025	10.0	0.359
13	59	6101	Parietal_Sup_L	0.024	28.5	0.383
14	9	2211	Frontal_Mid_Orb_L	0.024	86.5	0.449
15	86	8202	Temporal_Mid_R	0.024	60.5	0.419
16	13	2311	Frontal_Inf_Tri_L	0.023	3.5	0.341
17	94	9012	Cerebelum_Crus2_R	0.022	60.5	0.419
18	25	2611	Frontal_Med_Orb_L	0.022	86.5	0.449
19	67	6301	Precuneus_L	0.021	28.5	0.383
20	29	3001	Insula_L	0.019	53.5	0.413
21	110	9110	Vermis_3	0.018	79.0	0.443
22	84	8122	Temporal_Pole_Sup_R	0.018	13.5	0.365
23	15	2321	Frontal_Inf_Orb_L	0.018	79.0	0.443
24	70	6402	Paracentral_Lobule_R	0.017	86.5	0.449
25	55	5401	Fusiform_L	0.017	79.0	0.443
26	14	2312	Frontal_Inf_Tri_R	0.016	111.0	0.509
27	47	5021	Lingual_L	0.016	53.5	0.413
28	92	9002	Cerebelum_Crus1_R	0.015	22.0	0.377
29	102	9052	Cerebelum_7b_R	0.014	40.5	0.401
30	26	2612	Frontal_Med_Orb_R	0.014	60.5	0.419
31	4	2102	Frontal_Sup_R	0.013	22.0	0.377
32	77	7101	Thalamus_L	0.012	40.5	0.401
33	34	4012	Cingulum_Mid_R	0.012	28.5	0.383
34	39	4111	ParaHippocampal_L	0.012	60.5	0.419
35	57	6001	Postcentral_L	0.011	28.5	0.383
36	104	9062	Cerebelum_8_R	0.011	17.0	0.371
37	91	9001	Cerebelum_Crus1_L	0.011	10.0	0.359
38	66	6222	Angular_R	0.010	99.0	0.479
39	74	7012	Putamen_R	0.010	47.0	0.407
40	16	2322	Frontal_Inf_Orb_R	0.009	60.5	0.419
41	32	4002	Cingulum_Ant_R	0.008	66.5	0.431
42	20	2402	Supp_Motor_Area_R	0.008	53.5	0.413
43	101	9051	Cerebelum_7b_L	0.008	40.5	0.401
44	75	7021	Pallidum_L	0.008	114.5	0.521
45	7	2201	Frontal_Mid_L	0.008	28.5	0.383
46	58	6002	Postcentral_R	0.007	1.0	0.299
47	27	2701	Rectus_L	0.007	71.5	0.437
48	111	9120	Vermis_4_5	0.007	95.0	0.467
49	62	6202	Parietal_Inf_R	0.007	35.5	0.395
50	68	6302	Precuneus_R	0.007	53.5	0.413
51	85	8201	Temporal_Mid_L	0.007	3.5	0.341
52	78	7102	Thalamus_R	0.007	40.5	0.401
53	54	5302	Occipital_Inf_R	0.006	71.5	0.437
54	22	2502	Olfactory_R	0.005	71.5	0.437
55	24	2602	Frontal_Sup_Medial_R	0.005	17.0	0.371
56	17	2331	Rolandic_Oper_L	0.005	13.5	0.365
57	8	2202	Frontal_Mid_R	0.005	13.5	0.365
58	49	5101	Occipital_Sup_L	0.003	91.0	0.455
59	28	2702	Rectus_R	0.003	10.0	0.359
60	61	6201	Parietal_Inf_L	0.003	40.5	0.401
61	21	2501	Olfactory_L	0.003	86.5	0.449
62	3	2101	Frontal_Sup_L	0.002	22.0	0.377
63	10	2212	Frontal_Mid_Orb_R	0.002	47.0	0.407
64	97	9031	Cerebelum_4_5_L	0.001	66.5	0.431
65	114	9150	Vermis_8	0.001	108.5	0.503
66	100	9042	Cerebelum_6_R	0.000	86.5	0.449

Table 3.15: A result summary of relative importance of different brain regions for ABIDE analysis based on Experiment III

LPBoost					SVM	
Rank	Region #	Region Code	Region Name	Coefficient	Rank	PE
1	86	8202	Temporal_Mid_R	1.000	39.0	0.379

CHAPTER IV

Topic 3: Bayesian Nonparameteric Inference on Peaks via Spatially Adaptive Non-stationary Gaussian Processes

4.1 Introduction

There has been an increasing interest in making inference on peaks of one-dimensional curves and multi-dimensional surfaces, motivated by a very broad range of biological and biomedical research topics. For example, in proteomics, mass spectrometry (MS) data have been collected for protein identification and quantifications, which poses challenge to identify peak locations and magnitudes in the MS data (*Dass and Brodbelt*, 2001; *House et al.*, 2011). In metabolomics, there are growing interests in analyzing the coupling of liquid chromatography and mass spectrometry (LC-MS) data to facilitate metabolite identification and quantitation, which requires peak detections on noisy data with unknown smoothness (*Zhou et al.*, 2012). In genomics, it has been paid attentions by biologists in detecting the feature (e.g. local extrema) of expression profiles in the cDNA microarray experiments (*Raghuraman et al.*, 2001; *Song et al.*, 2006). In the research of hormonal disruptions during the reproductive cycle, surge times and magnitudes of hormone trajectories are considered important features to make inference (*Veiga-Lopez et al.*, 2008; *Kang et al.*, 2012). In neuroimaging study, identifying peak location also relate to many implications. For example, using brain electroencephalogram (EEG) data to detect the peak location and magnitude of brain activity after certain stimulus in a temporal domain is capable of assessing regular brain functions so that assist clinical diagnosis of disease (*Liu et al.*, 2013; *Rangaswamy and Porjesz*, 2014). It is also of great interest to localize the brain activities by detecting the peak activation locations in a three-dimensional brain in order to study particular brain functions or diseases (*Lindquist*, 2008; *Kang et al.*, 2011).

Several methods have been proposed in the literature for identifying peaks locations from different biological applications. For example, *Raghuraman et al.* (2001) used moving-average smoothing and Fourior convolution smoothing in the application to microarray data anaysis. *Song et al.* (2006) proposed a non-parametric kernel smoothing technique that enables statistical inference on peak locations and applied the method to microarray dataset. *Kang et al.* (2012) proposed a local kernel smoothing method that further utilize analysis on multiple curves using the non-parametric mixed-effects model. Wavelet regression method has also been used to model peaks, in mass spectrometry study (*Morris et al.*, 2008), however, the scales and locations that index the wavelet basis functions have difficulties to relate with any biological interpretation. *House et al.* (2011) presented a nonparametric Bayesian approach based on Levy Adaptive Regression Kernels to model mass spectral data. Their methods are directly interpretable and reply on informative prior distributions to define the peak resolution. *Kang et al.* (2011) provided a hierarchical spatial point process model to locate peak activation centers.

In this study, we aim to build a peak model via Gaussian process regression. We utilize the good properties of Gaussian process in derivative calculation. The Karhunen-Loéve (K-L) expansion of Gaussian process techniques are also adopted to enable high dimensional computation. In simulation study, we demonstrate the good property of proposed method compared to the nonparametric kernel smoothing method proposed by *Song et al.* (2006). In real data analysis, we demonstrate the use of proposed method to find peak locations of one dimensional EEG data on temporal domain, and also to locate the peak of power on the frequency domain. Based on the peaks and the corresponding magnitude captured by the proposed methods, we investigate the influence of alcohol on brain functions and also build a classifier based on EEG measurement to predict the use of alcohol.

4.2 Methods

Suppose we collect data from n independent observations, denoted $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, where $y_i \in \mathbb{R}$ is the outcome variable and $\mathbf{x}_i \in \mathbb{R}^d$ predictors. We assume that

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \tag{4.1}$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ and $f : \mathbb{R}^d \to \mathbb{R}$ is a real-valued function defined on $\mathbb{R}^d$. We model funcitontion f as one realization of the gaussian process with mean function μ and covariance kernel κ, denoted

$$f \sim \mathcal{GP}(\mu, \kappa). \tag{4.2}$$

We consider the Karhunen-Loéve (KL) expansion on f, which is given by

$$f(\mathbf{x}) = \sum_{l=0}^{\infty} \theta_l \psi_l(\mathbf{x}), \text{ with } \theta_l \sim \mathrm{N}(0, \lambda_l), \tag{4.3}$$

where $\{\lambda_l\}_{l=0}^{\infty}$ and $\{\psi_l(\mathbf{x})\}_{l=0}^{\infty}$ are the eigenvalues and eigenfunctions of the covariance kernel $\kappa(\mathbf{x}, \mathbf{x}')$, i.e. they are satisfied with the eigen equations.

$$\int\limits_{\mathbb{R}^d} \psi_l(\mathbf{x})\kappa(\mathbf{x}, \mathbf{x}')\mathrm{d}\mathbf{x} = \lambda_l \psi_l(\mathbf{x}'), \tag{4.4}$$

where $\{\psi_l(\mathbf{x})\}_{l=0}^{\infty}$ are orthogonal functions over $\mathbb{R}^d$, i.e.,

$$\int\limits_{\mathbb{R}^d} \psi_l(\mathbf{x})\psi_{l'}(\mathbf{x})\mathrm{d}\mathbf{x} = 0, \qquad \text{and} \qquad \int\limits_{\mathbb{R}^d} \psi_l^2(\mathbf{x})\mathrm{d}\mathbf{x} = 1.$$

The peaks and magnitudes of a function can be naturally determined by its derivatives for a one-dimensional case or the gradient and the Hessian matrix for a multi-dimensional case. The gradient and Hessian matrix of a function $g(\mathbf{x})$ are respectively defined as

$$\nabla g = \left(\frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_d}\right)^{\mathrm{T}}, \quad \mathrm{H}g = \begin{bmatrix} \frac{\partial^2 g}{\partial x_1^2} & \frac{\partial^2 g}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 g}{\partial x_1 \partial x_d} \\ \frac{\partial^2 g}{\partial x_2 \partial x_1} & \frac{\partial^2 g}{\partial x_2^2} & \cdots & \frac{\partial^2 g}{\partial x_2 \partial x_d} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 g}{\partial x_d \partial x_1} & \frac{\partial^2 g}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 g}{\partial x_d^2} \end{bmatrix}. \tag{4.5}$$

By the linear property of the derivatives, we have

$$\nabla f(\mathbf{x}) = \sum_{l=0}^{\infty} \theta_l \nabla \psi_l(\mathbf{x}), \quad \text{and} \quad \mathrm{H}f(\mathbf{x}) = \sum_{l=0}^{\infty} \theta_l \mathrm{H}\psi_l(\mathbf{x}). \tag{4.6}$$

Next, we consider the truncation of the K-L expansion for approximations on $f(\mathbf{x})$, $\nabla f(\mathbf{x})$ and $\mathrm{H}f(\mathbf{x})$, for a sufficient large integer L, $f(\mathbf{x})$ can be approximately represented as

$$f(\mathbf{x}) = \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\theta}, \quad \text{and} \quad \boldsymbol{\theta} \sim \mathrm{N}(\mathbf{0}_{L+1}, \boldsymbol{\Lambda}), \tag{4.7}$$

where $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_L)^{\mathrm{T}}$, $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_0, \ldots, \lambda_L\}$ and $\boldsymbol{\psi}(\mathbf{x}) = [\psi_0(\mathbf{x}), \ldots, \psi_L(\mathbf{x})]^{\mathrm{T}}$. Furthermore, define $\nabla\boldsymbol{\psi}(\mathbf{x}) = [\nabla\psi_0(\mathbf{x}), \ldots, \nabla\psi_L(\mathbf{x})]$ of dimension $d \times (L+1)$ and $\mathrm{H}\boldsymbol{\psi}(\mathbf{x}) = [\mathrm{H}\psi_0(\mathbf{x}), \ldots, \mathrm{H}\psi_L(\mathbf{x})]$ of dimension $d \times (L+1)d$. Then we have

$$\nabla f(\mathbf{x}) = \nabla\boldsymbol{\psi}(\mathbf{x})\boldsymbol{\theta}, \quad \text{and} \quad \mathrm{H}f(\mathbf{x}) = \mathrm{H}\boldsymbol{\psi}(\mathbf{x})(\boldsymbol{\theta} \otimes \mathbf{I}_d), \tag{4.8}$$

where $\mathbf{I}_d$ is a $d \times d$ identity matrix and "$\otimes$" represents the Kronecker product. The sampling distribution of y_i is given by

$$[y_i \mid \mathbf{x}_i, \boldsymbol{\theta}\ ] \sim \mathrm{N}(\boldsymbol{\psi}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\theta}, \sigma^2)$$

Let $\mathbf{z}$ and η respectively represent the location and magnitude of one peak of the function $f(\mathbf{x})$ on $\mathbb{R}^d$. Then given $\mathbf{z}$ and η, the sampling distribution of function f can be determined the distribution of $\boldsymbol{\theta}$, which is a truncated normal distribution given by

$$[\boldsymbol{\theta} \mid \mathbf{z}, \eta] \sim \mathrm{T}N_{\mathcal{A}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}], \tag{4.9}$$

where $\boldsymbol{\mu} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\binom{\eta}{0}$, $\boldsymbol{\Sigma} = \boldsymbol{\Lambda} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^{\mathrm{T}}$ and $\mathcal{A} = \{\boldsymbol{\theta} : \mathrm{H}\boldsymbol{\psi}(\mathbf{z})(\boldsymbol{\theta} \otimes \mathbf{I}_d) < 0\}$ with

$$\boldsymbol{\Sigma}_{12} = \begin{pmatrix} \boldsymbol{\psi}^{\mathrm{T}}(\mathbf{z})\boldsymbol{\Lambda} \\ \nabla\boldsymbol{\psi}(\mathbf{z})\boldsymbol{\Lambda} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\Sigma}_{22} = \begin{pmatrix} \boldsymbol{\psi}^{\mathrm{T}}(\mathbf{z})\boldsymbol{\Lambda}\boldsymbol{\psi}(\mathbf{z}) & \boldsymbol{\psi}^{\mathrm{T}}(\mathbf{z})\boldsymbol{\Lambda}\nabla\boldsymbol{\psi}^{\mathrm{T}}(\mathbf{z}) \\ \nabla\boldsymbol{\psi}(\mathbf{z})\boldsymbol{\Lambda}\boldsymbol{\psi}(\mathbf{z}) & \nabla\boldsymbol{\psi}(\mathbf{z})\boldsymbol{\Lambda}\nabla\boldsymbol{\psi}^{\mathrm{T}}(\mathbf{z}) \end{pmatrix}. \tag{4.10}$$

4.2.1 Posterior Computation

Suppose we assign joint priors on the location $\mathbf{z}$ and the magnitude η, denoted $\pi(\mathbf{z}, \eta)$. The joint posterior distribution of $\mathbf{z}$, η and $\boldsymbol{\theta}$ is given by

$$\pi(\mathbf{z}, \eta, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{x}) \propto \pi(\mathbf{z}, \eta)\pi(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{x}, \mathbf{z}, \eta). \tag{4.11}$$

This is the target distribution of the posterior simulation. The full conditional of $\boldsymbol{\theta}$ given other parameters is a truncated normal distribution

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{x}, \mathbf{z}, \eta) \propto \prod_{i=1}^{n} \pi(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{z}, \eta).$$

This implies that

$$[\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{x}, \mathbf{z}, \eta] \propto \mathrm{T}N_{\mathcal{A}}(\boldsymbol{\nu}, \mathbf{S}), \tag{4.12}$$

where the truncation area $\mathcal{A}$ is defined the same as in (4.9). The mean and covariance matrix is given by

$$\boldsymbol{\nu} = \mathbf{S}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i \boldsymbol{\psi}^{\mathrm{T}}(\mathbf{x}_i) + \boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}\right], \quad \text{and} \quad \mathbf{S} = \left[\frac{1}{\sigma^2}\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{x}_i)\boldsymbol{\psi}^{\mathrm{T}}(\mathbf{x}_i) + \boldsymbol{\Sigma}^{-1}\right]^{-1}.$$

Let $g(\mathbf{z}, \eta)$ be a joint proposal function for peak location $\mathbf{z}$ and peak magnitude η. We have the following independent Metropolis-Hastings algorithm to perform the posterior computation. Given a set of initial values of $\{\boldsymbol{\theta}^{(0)}, \mathbf{z}^{(0)}, \eta^{(0)}\}$, in the kth iteration, for $k = 1, \ldots, K$,

- Draw $(\mathbf{z}^*, \eta^*) \sim g(\mathbf{z}, \eta)$ and given $(\mathbf{z}^*, \eta^*)$, draw $\boldsymbol{\theta}^*$ from (4.12).

- Compute the acceptance ratio

$$\alpha = \min\left\{\frac{\pi(\mathbf{z}^*, \eta^*) g(\mathbf{z}^{(k-1)}, \eta^{(k-1)})}{\pi(\mathbf{z}^{(k-1)}, \eta^{(k-1)}) g(\mathbf{z}^*, \eta^*)}\right\}$$

- Draw $u \sim \mathrm{U}(0,1)$, set $(\mathbf{z}^{(k)}, \eta^{(k)}, \boldsymbol{\theta}^{(k)}) = (\mathbf{z}^*, \eta^*, \boldsymbol{\theta}^*)$ if $u < \alpha$, set $(\mathbf{z}^{(k)}, \eta^{(k)}, \boldsymbol{\theta}^{(k)}) = (\mathbf{z}^{(k-1)}, \eta^{(k-1)}, \boldsymbol{\theta}^{(k-1)})$, otherwise.

Of note, if we let $g(\mathbf{z}, \eta) = \pi(\mathbf{z}, \eta)$ then we always have $\alpha = 1$.

4.3 Simulation Study

In the following, a simulation study is performed to examine the performance of our proposed model. The power and robustness of the proposed method is evaluated under the effects of different noise level and of the existence of multiple equilibrium points. We compare our methods with the nonparametric kernel smoothing (NKS) method proposed by*Song et al.* (2006) in regards to both point and interval estimation for the location of local extrema.

4.3.1 Effect of Noise

The first simulation focus on the effects of noise on the robustness of the proposed method. The data are generated from the following model with a single extrema point at coordinate $x_0 = -0.6$.

$$y_i = f_1(x_i) + \epsilon_i,\ i = 1, \ldots, 300, \tag{4.13}$$

$$f_1(x) = 0.01 \times \exp(-10 \times (x + 0.6)^2 + 5) \tag{4.14}$$

where ϵ_i s were independently simulated from $\mathcal{N}(0, \sigma^2)$. Clearly, the peak location $x_0 = -0.6$. The performance of the proposed method was assessed at different variance values, each based on 100 replicates. Table (4.1) presents the results.

Table 4.1: Simulation results for the proposed method compared to the nonparametric kernel smoothing (NKS) method in regards to mean squared error (MSE) of peak location and 95% confidence interval coverage

	Proposed Method		NKS	
	MSE ($\times 10^4$)	Coverage	MSE ($\times 10^4$)	Coverage
sd = 0.1	0.4752	1.00	1.5495	0.98
sd = 0.3	2.5390	1.00	7.5331	1.00
sd = 0.5	3.5605	1.00	10.3270	0.99

4.3.2 Effect of Multiple Equilibrium Points

The second simulation focus on the performance of the proposed peak identifier on locating multiple peaks. We simulate a function with two peaks of different smoothness.

$$y_i = f_2(x_i) + \epsilon_i, \; i = 1, \ldots, 300, \tag{4.15}$$

$$f_2(x) = \begin{cases} \sin(2\pi x) & \text{if } x < 0 \\ -10 \times (x - 0.5)^2 + 2.5 & \text{if } x \geq 0 \end{cases} \tag{4.16}$$

Clearly, two peaks locate at $x_0^{(1)} = -0.75$ and $x_0^{(2)} = 0.5$. Again, ϵ_i s were independently simulated from $\mathcal{N}(0, \sigma^2)$. Based on 100 replications, we summarize the results in Table (4.2).

Table 4.2: Simulation results for proposed method compared to NKS method in regards to MSE of peak location and 95% confidence interval coverage for two peak locations respectively

	Proposed Method		NKS	
	MSE ($\times 10^4$)	coverage	MSE ($\times 10^4$)	coverage
$x_0^{(1)} = -0.75$				
sd = 0.1	1.42	1	2.00	0.92
sd = 0.3	2.84	0.99	4.47	0.91
sd = 0.5	3.23	0.98	10.07	0.92
$x_0^{(2)} = 0.5$				
sd = 0.1	1.13	1	160.25	0.93
sd = 0.3	1.97	1	20.54	0.83
sd = 0.5	4.28	0.99	318.95	0.88

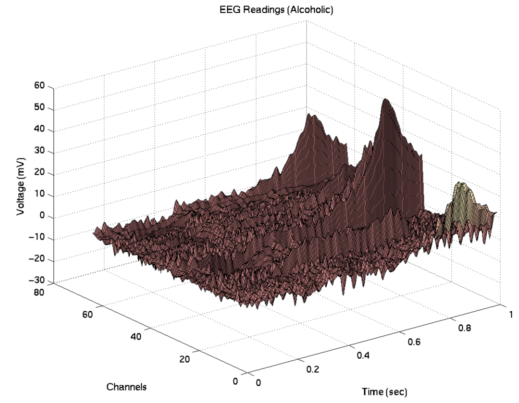
4.4 EEG Data Analysis

The methods are applied to the data from an electroencephalography (EEG) study of alcoholism; see `http://kdd.ics.uci.edu/datasets/eeg/eeg.data.html`. The objective of our analysis is to estimate the relationship between alcoholism and brain activity through peak location and magnitude. The study compromises 122 subjects: 77 alcoholic subjects and 45 non-alcoholic controls. For each subject, 64 electrodes were placed on their scalp and EEG was recorded from each electrode at a frequency of 256 Hz (3.9-msec epoch) for 1 second. Each subject was exposed to either a single stimulus (S1) or to two stimuli (S1 and S2) which were pictures of objects chosen from the 1980 Snodgrass and Vanderwart picture set. When two stimuli were shown, they were presented in either a matched condition where S1 was identical to S2 or in a non-matched condition where S1 differed from S2. For simplicity, we name Experiment I, II and III representing each stimulus design. The number of trials conducted using three experiments for each subject is shown in Table (4.8) and (4.9) in the Appendix. We consider the average EEG across all trials under the same experiment. Note that there are 17 trials with empty files for subject co2c1000367 and are excluded from taking average. Subject co2c0000392 doesn't have any record under electrodes "X". As a result we remove this subject from analysis.

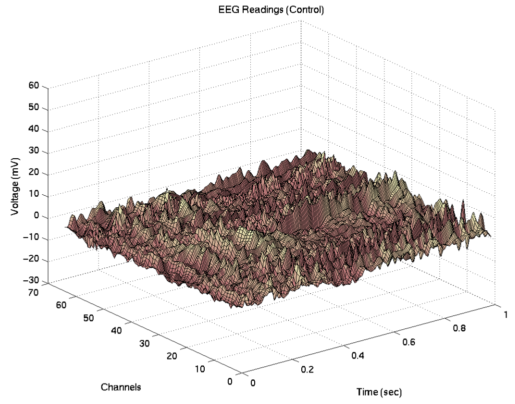
Experiment I	Each subject was exposed to either a single stimulus (S1)
Experiment II	Each subject was exposed to two identical stimuli (S1 and S2 are identical)
Experiment III	Each subject was exposed to two non-matched stimuli (S1 and S2 are different)

The data have been previously described and analyzed in *Li et al.* (2010a), *Hung and Wang* (2012), *Zhou and Li* (2014), and *Kang et al.* (2016). *Zhang et al.* (1995) described in detail the data collection process. Figure (4.4) shows example plots of a control and alcoholic subject. The plots indicate voltage, time, and channel and are averaged over 10 trials for the single stimulus condition. The electrode positions were located at standard sites (Standard Electrode Position Nomenclature, American Electroencephalographic Association 1990). The spatial structure of the electrodes is shown in Figure (4.4), which was recovered from the standard electrode position nomenclature described by Fig. 1 of `https://www.acns.org/pdf/guidelines/Guideline-5.pdf`.

Alcohol

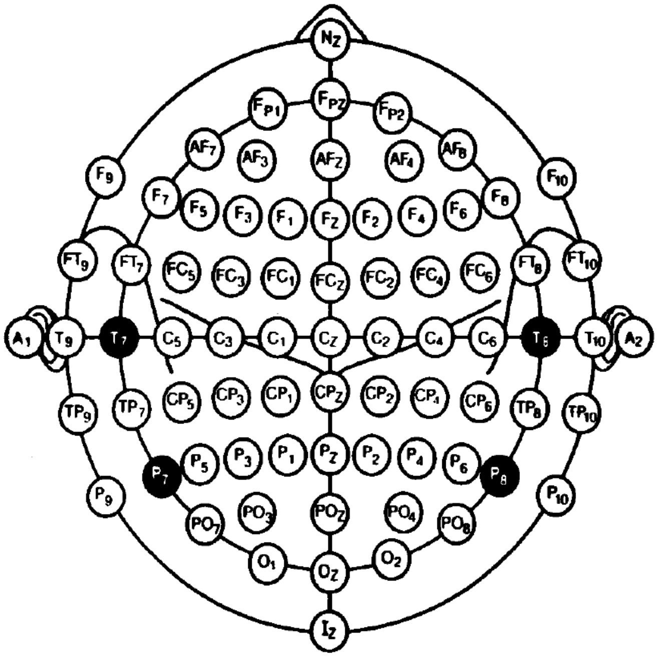


Control



Alpha band, the frequency of which range from 8 Hz to 14 Hz, has been described as an index of relaxed wakefulness and closely related to subjects' alcoholic status. Research has shown that EEG peak alpha frequency (PAF) (measured in Hz) has been correlated to cognitive performance between healthy and clinical individuals, and among healthy individuals. PAF also varies within individuals across developmental stages, among different cognitive tasks, and among physiological states induced by administration of various substances. The present study suggests that, among other things, PAF reflects a trait or state of cognitive preparedness. Research has shown the prominent effects of low doses of alcohol include increases in slow alpha activity or lowering of alpha peak frequency, while moderate doses show increases in slow alpha and theta bands (Ehlers et al., 2004). Alpha activity has also been positively associated with desire to drink. We are considering to analyze the alpha band of EEG and the corresponding power of frequency that imply the alcoholic status.

In the following, we apply our proposed method to identify peak location and magnitude in 1) alpha band wave along one second time interval; 2) power of the frequency curve on 0 to 20 Hz frequency domain; and 3) build classification model using the results from 1) and 2) to distinguish alcoholic subjects from normal control, and to study the difference in brain activity between the two groups under three experiments.



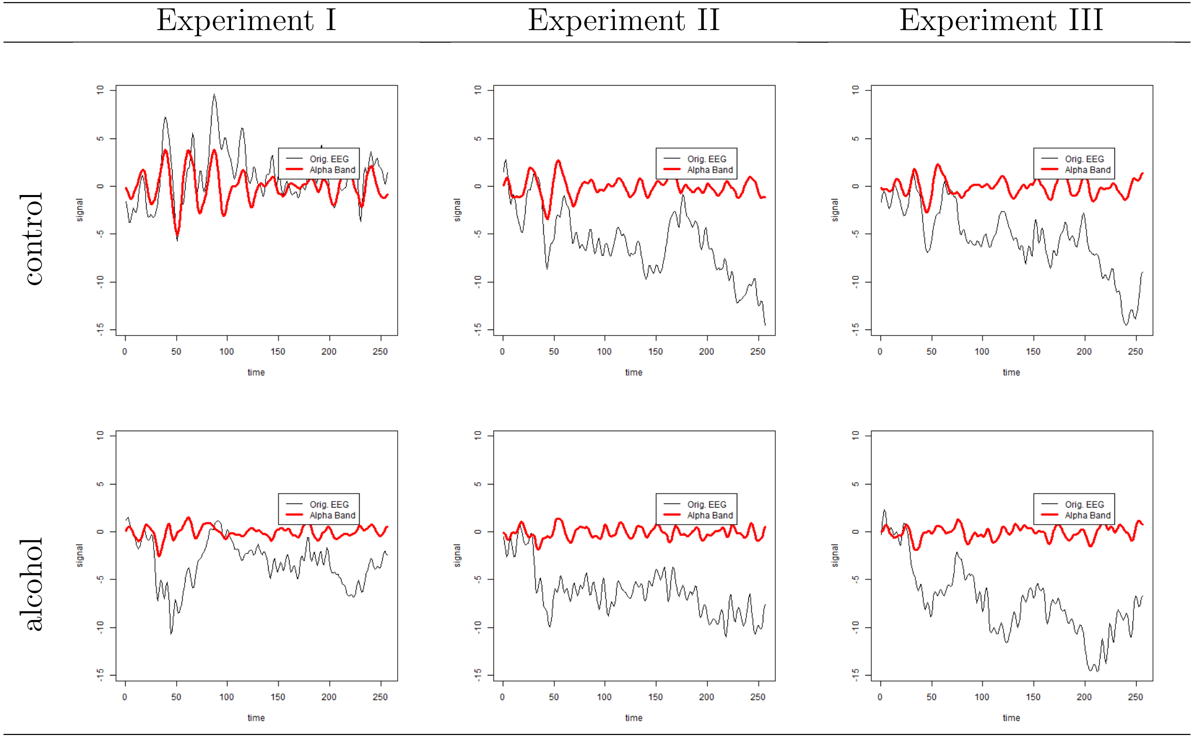
4.4.1 Peak of Alpha Band in Temporal Domain

The EEG is typically described in terms of rhythmic activity that can be divided into bands by frequency according to different wave patterns. The frequency range has a certain distribution over the scalp or a certain biological significance. Most of the cerebral signal observed in the scalp EEG falls in the range of 120 Hz. The common EEG frequency band include:

Delta	< 4
Theta	≥ 4 and < 8
Alpha	≥ 8 and < 14
Beta	$\geq$14

Alpha band, the frequency of which range from 8 Hz to 14 Hz, merges with closing of the eyes and relaxation and attenuates with eye opening or mental exertion. It is considered as the posterior basic rhythm and is usually seen in the posterior regions of the head on both sides, higher in amplitude on the dominant side. It is described as an index of relaxed wakefulness. Alpha is slower in young children (closer to theta frequency) and increases with age into high alpha frequencies and is a key feature of EEG maturation (Niedermayer and Lopes Da Silva, 1999); alpha power is stable throughout adult life. In order to achieve alpha band from the original EEG signal, we use the Butterworth square-wave filter described in *Pollock* (2000) to filter out waves that have frequency lower than 8 Hz and higher than 14 Hz. After Butterworth filter, the original EEG data, plotted in Figure (4.4), was transformed to alpha band curves, plotted in Figure (4.5).

Figure 4.1: Plots of original EEG signal and the filtered alpha wave on electrode P6 of subject "co2c0000337" (first row) from control group and subject "co2a0000364" from alcohol group. Data collected from different stimulus types are shown in three columns



4.4.2 Peak of Power in Frequency Domain

In this section, the peak of alpha frequency is examined for difference between control group and the group with alcohol usage. Table (4.6) lists the electrode locations that the peak alpha frequency is significantly different among two groups.

Figure 4.2: Plots of the fitted curve (blue line) for the filtered alpha wave and identified peak location (red triangle) by the proposed method. The example data is the alpha wave from electrode CP2 of subject ”co2c0000337” (first row) from control group and subject ”co2a0000364” from alcohol group and across different stimulus types

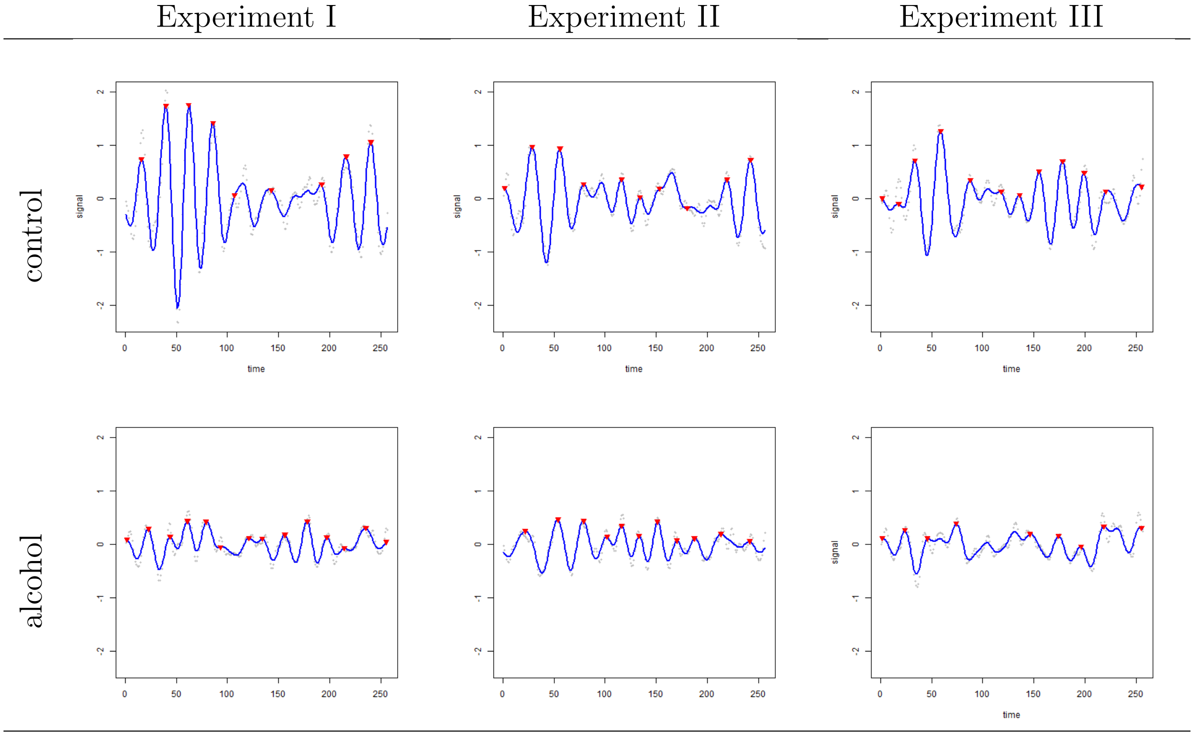


Figure (4.3) shows peak locations identified by the proposed method for P6 for one subject in the control group and one subject in the alcohol group.

4.4.3 Classification based on Peak and Magnitude

In this section, the peak locations and magnitude identified using the proposed method are used as predictors to predict a subject’s alcoholic status. Table (4.7) shows the classification error calculated given different peak location predictors, including first peak location, global peak location, and average peak interval in time domain, and peak alpha frequency in frequency domain. We achieve the overall accuracy of 20%.

Table 4.3: A list of electrodes that show significant difference in time locations of the first peak after stimulus on alpha wave. The hypothesis test is based on the Student's T test and exams the quantity of interest between control group and alcohol group under three different experiments

Location	Control Group(1-256)	Alcohol Group((1-256)	p value
Experiment I			
T7	97.80	124.39	0.0443
P7	61.96	81.08	0.0492
Experiment II			
FZ	122.31	91.57	0.0177
C5	99.22	128.57	0.0407
F2	123.98	98.13	0.0453
Experiment III			
AF2	127.78	88.42	0.0035
FZ	115.20	89.47	0.0320
FPZ	142.51	106.36	0.0109
AFZ	126.36	93.81	0.0125
FCZ	119.80	92.47	0.0399

Table 4.4: A list of electrodes that show significant difference in time locations of the highest peak after stimulus on alpha wave. The hypothesis test is based on the Student's T test and exams the quantity of interest between control group and alcohol group under three different experiments

Location	Control Group(1-256)	Alcohol Group((1-256)	p value
Experiment I			
C3	10.07	7.22	0.0108
C4	11.60	8.17	0.0133
C2	10.76	7.53	0.0241
PO7	10.42	7.57	0.0248
OZ	10.87	7.60	0.0122
Experiment II			
O1	8.58	6.32	0.0263
FT7	5.96	8.00	0.0416
PO7	8.42	6.57	0.0401
Experiment III			
CP2	6.82	8.84	0.0372
P3	6.58	8.62	0.0316
P4	6.58	9.96	0.0007
PZ	6.31	8.92	0.0028
C2	6.16	8.08	0.0468
PO8	6.02	8.18	0.0124
P2	6.36	9.22	0.0017
P1	6.82	8.64	0.0443

Table 4.5: A list of electrodes that show significant difference in average interval of peak locations on alpha wave. The hypothesis test is based on the Student's T test and exams the quantity of interest between control group and alcohol group under three different experiments

Location	Control Group(1-256)	Alcohol Group((1-256)	p value
Experiment I			
FP1	22.69	21.83	0.0278
FP2	22.56	21.65	0.0185
F7	22.47	21.46	0.0162
AF2	22.77	21.85	0.0196
FZ	22.54	21.78	0.0403
F4	22.37	21.40	0.0305
FC5	22.40	21.54	0.0285
FC2	22.23	21.32	0.0173
T8	22.27	21.27	0.0234
C4	22.28	21.37	0.0212
CP5	22.98	22.08	0.0277
CP2	22.84	21.76	0.0070
X	22.71	21.73	0.0232
F6	22.16	21.21	0.0217
FC3	22.37	21.45	0.0102
TP8	22.56	21.55	0.0262
TP7	22.85	21.96	0.0446
AFZ	22.80	21.80	0.0167
C1	22.55	21.66	0.0243
FCZ	22.78	21.91	0.0159
CPZ	22.94	22.07	0.0079
Experiment II			
FP1	22.55	21.60	0.0058
FP2	22.73	21.51	0.0009
F8	22.42	20.85	0.0001
FC6	22.47	20.86	0.0000
T8	22.19	20.68	0.0001
T7	22.43	21.53	0.0105
CZ	22.12	21.15	0.0101
C4	22.03	20.93	0.0018
CP5	22.42	21.64	0.0365
CP6	22.24	21.55	0.0461
P3	22.57	21.76	0.0175
X	22.36	21.52	0.0286
AF7	22.13	21.23	0.0277
AF8	22.32	21.25	0.0107
F5	22.12	21.36	0.0478
F6	22.17	21.21	0.0146
FT8	22.30	20.69	0.0001
FPZ	22.69	21.92	0.0423
C6	22.23	20.90	0.0001
TP8	22.39	21.09	0.0005
TP7	22.43	21.70	0.0481
CP4	22.08	21.38	0.0285
P5	22.62	21.84	0.0239
C2	22.11	21.04	0.0053
nd	22.41	21.60	0.0334
Experiment III			
F8	22.24	21.42	0.0409
FC5	22.61	21.64	0.0166
C3	22.43	21.62	0.0332
CP2	22.73	21.61	0.0031
P4	22.79	21.72	0.0046
PZ	23.00	21.70	0.0020
P8	23.10	21.79	0.0001
PO2	22.50	21.72	0.0366
PO1	22.76	21.68	0.0083
O2	22.56	21.66	0.0201
F6	22.36	21.29	0.0018
FC3	22.45	21.50	0.0066
C5	22.79	21.92	0.0299
TP8	22.59	21.78	0.0428
P6	22.74	21.86	0.0091
C2	22.10	21.34	0.0146
PO8	22.43	21.74	0.0418
FCZ	22.82	22.04	0.0313
POZ	22.59	21.71	0.0177
P2	22.79	21.84	0.0070
P1	22.93	21.89	0.0143
CPZ	23.01	22.02	0.0228

Table 4.6: A list of electrodes that show significant difference in peak alpha frequency. The hypothesis test is based on the Student's T test and exams the quantity of interest between control group and alcohol group under three different experiments

Location	Control Group(8-14)	Alcohol Group((8-14)	p value
Experiment I			
CZ	11.75	10.97	0.0337
FPZ	11.82	10.89	0.0185
Experiment II			
FP1	12.67	11.64	0.0082
AF1	12.57	11.77	0.0398
P4	11.44	12.55	0.0038
F2	12.74	11.37	0.0007
Experiment III			
F7	13.07	12.01	0.0060
AF2	12.61	11.60	0.0192
C3	13.27	11.88	0.0003
AF7	12.92	12.06	0.0324
F5	13.31	11.49	0.0000
FPZ	12.27	11.38	0.0284
FC3	12.66	11.71	0.0274

Figure 4.3: Plots of the fitted curve (blue line) for the power curve of alpha wave and identified peak location (red triangle) by the proposed method. The example data is the power of frequency of the alpha wave on electrode P6 of subject "co2c0000337" (first row) from control group and subject "co2a0000364" from alcohol group and across different stimulus types

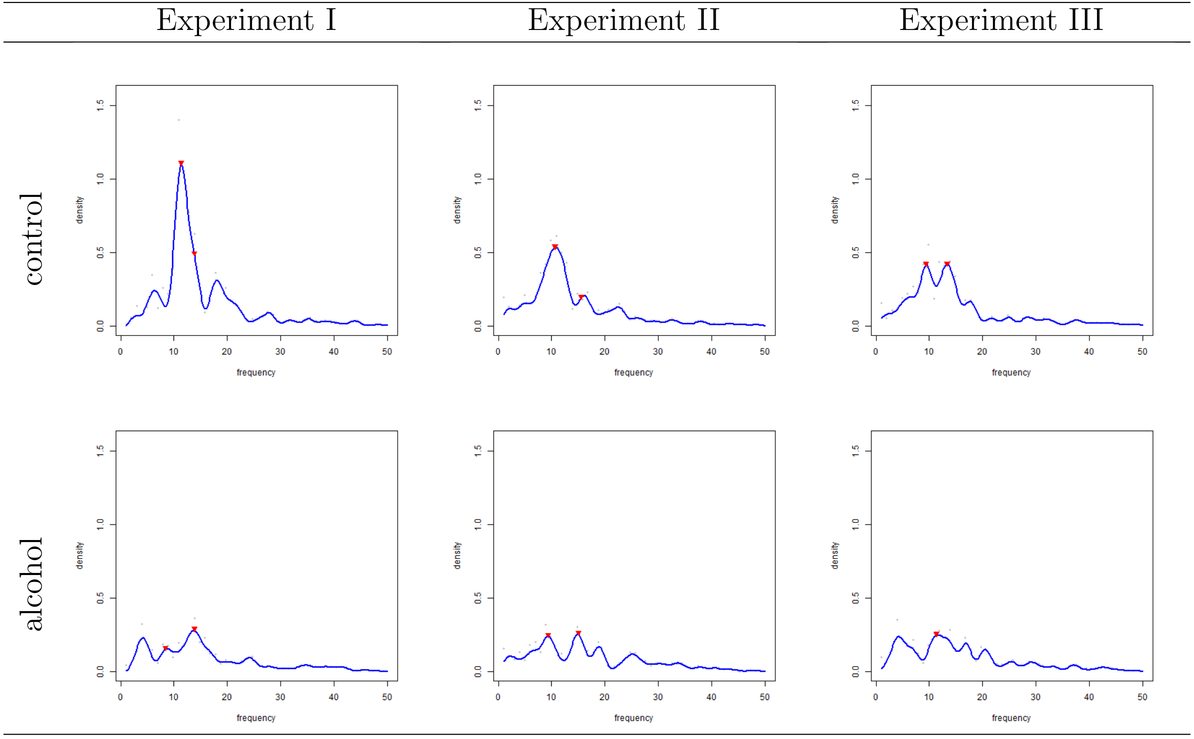


Table 4.7: EEG classification results based on three experiments presented in different rows and four types of location related predictors presented in different columns. The mean classification error and its corresponding standard error (in parenthesis) are calculated based on ten-fold cross validation

	First Loc.	Global Loc.	Ave. Interval	PAF	Combined All
Experiment I	0.33(0.05)	0.33(0.05)	0.36(0.05)	0.31(0.04)	0.29(0.02)
Experiment II	0.34(0.03)	0.29(0.03)	0.32(0.04)	0.24(0.05)	0.24(0.04)
Experiment III	0.31(0.04)	0.28(0.05)	0.36(0.04)	0.27(0.04)	0.25(0.03)
Combined (I, II, & III)	0.31(0.05)	0.28(0.05)	0.33(0.04)	0.26(0.04)	0.20(0.02)

4.5 Appendix

Figure 4.4: Original EEG signal curves for all control subjects in the first row and alcoholic subjects in the second row at electrode P6, averaged for all trials under each of the three experiments

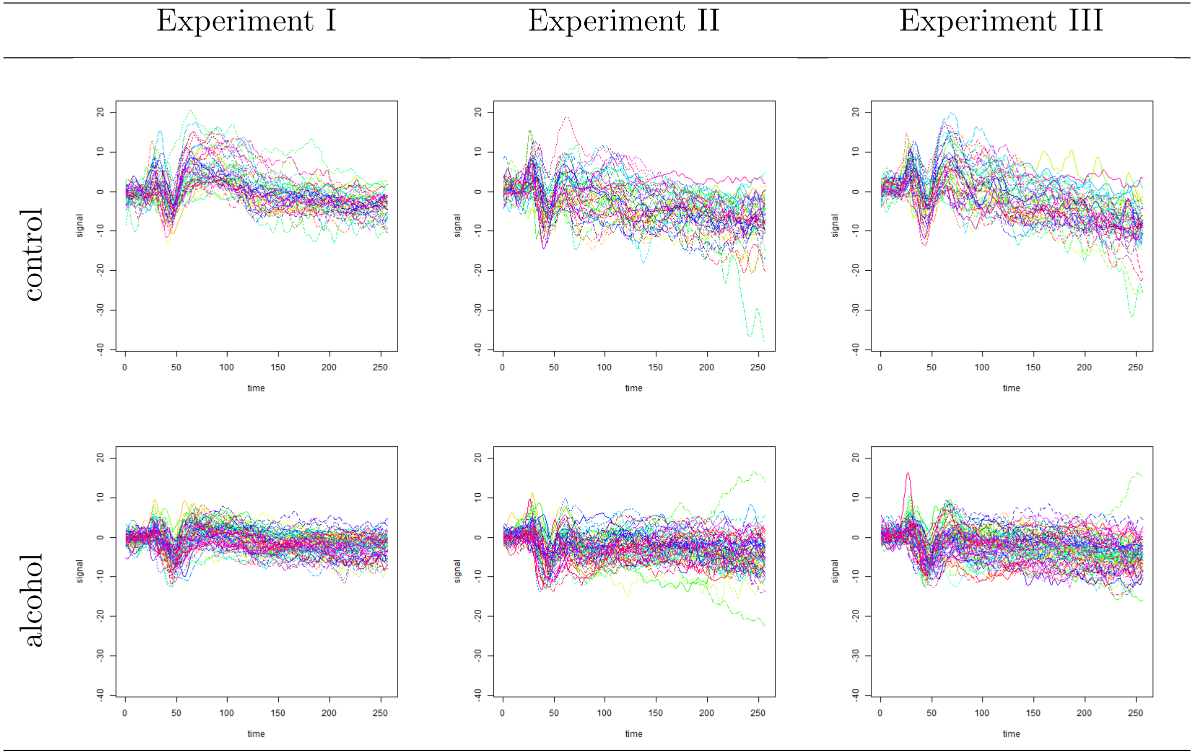


Figure 4.5: Filtered Alpha wave signals for all control subjects in the first row and alcoholic subjects in the second row at electrode P6, averaged across all trials under each of the three experiment

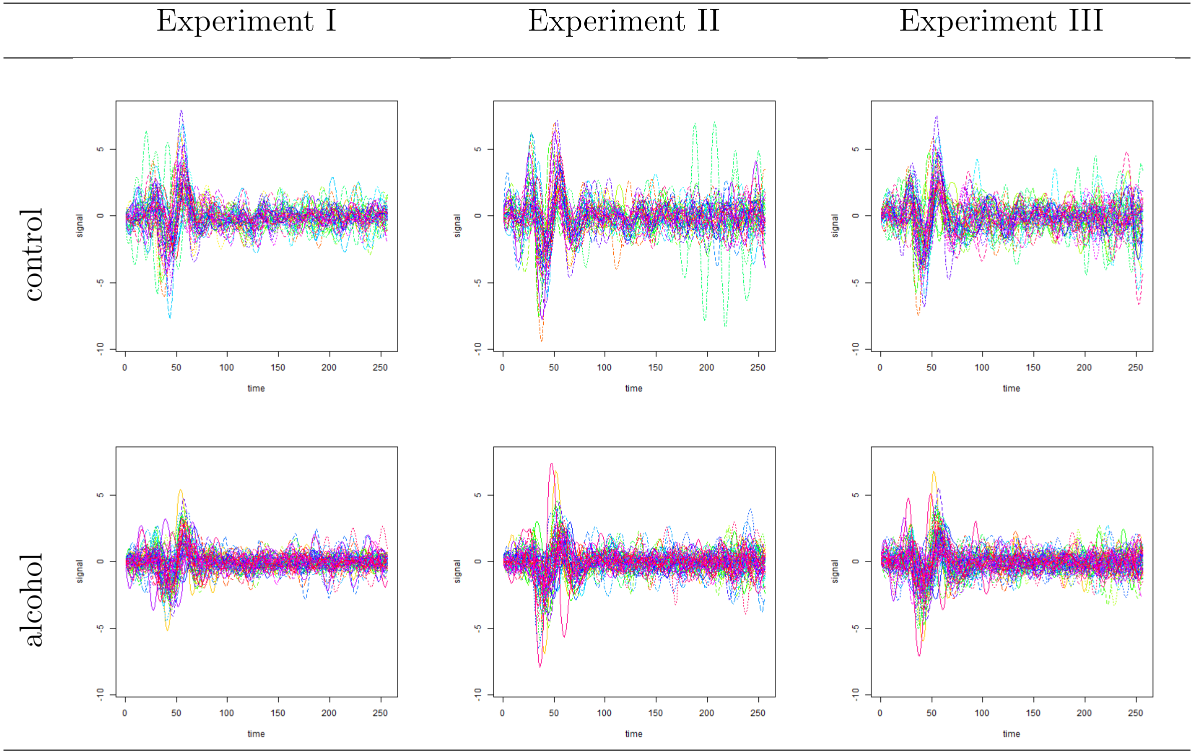


Figure 4.6: Alpha wave power curves for all control subjects in the first row and alcoholic subjects in the second row at electrode P6, averaged across all trials for each of the three experiments

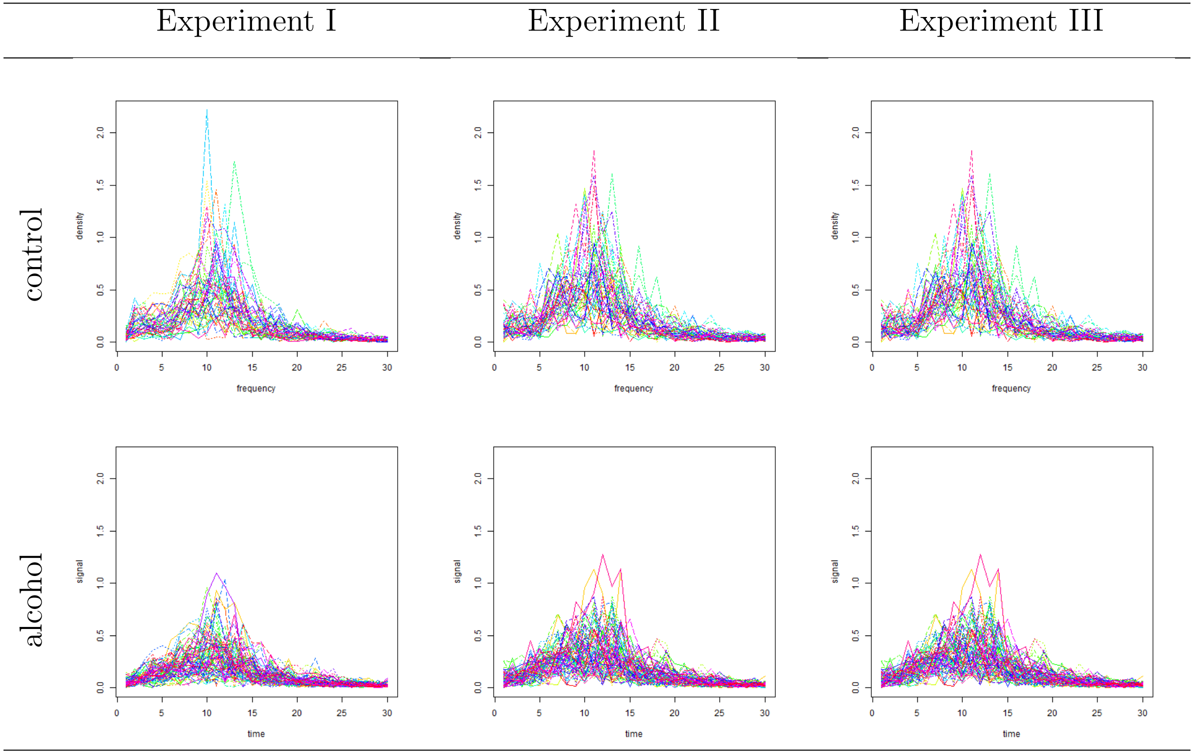


Table 4.8: A list of EEG data for alcoholism study in regards to the total number of trials of each subjects under three experiments

idx	subjects	# Trials with Experiment I	# Trials with Experiment II	# Trials with Experiment III	Total # of Trials
1	co2a0000364	40	27	21	88
2	co2a0000365	49	22	22	93
3	co2a0000368	39	14	20	73
4	co2a0000369	58	30	26	114
5	co2a0000370	59	29	29	117
6	co2a0000371	53	25	28	106
7	co2a0000372	57	29	29	115
8	co2a0000375	39	24	22	85
9	co2a0000377	50	25	25	100
10	co2a0000378	56	27	29	112
11	co2a0000379	54	27	26	107
12	co2a0000380	43	23	21	87
13	co2a0000381	54	25	23	102
14	co2a0000382	58	29	26	113
15	co2a0000384	49	22	29	100
16	co2a0000385	39	17	14	70
17	co2a0000386	35	20	17	72
18	co2a0000387	34	16	17	67
19	co2a0000388	38	25	24	87
20	co2a0000390	51	27	26	104
21	co2a0000392	42	28	26	96
22	co2a0000394	55	26	30	111
23	co2a0000395	46	24	28	98
24	co2a0000396	42	14	21	77
25	co2a0000398	37	30	29	96
26	co2a0000400	40	23	15	78
27	co2a0000402	50	26	16	92
28	co2a0000403	46	27	24	97
29	co2a0000404	54	25	16	95
30	co2a0000405	42	13	14	69
31	co2a0000406	30	17	13	60
32	co2a0000407	47	28	18	93
33	co2a0000409	52	26	28	106
34	co2a0000410	56	27	24	107
35	co2a0000411	48	23	23	94
36	co2a0000412	52	17	20	89
37	co2a0000414	56	30	29	115
38	co2a0000415	50	15	14	79
39	co2a0000416	36	23	25	84
40	co2a0000417	42	30	27	99
41	co2a0000418	48	25	27	100
42	co2a0000419	57	30	29	116
43	co2a0000421	49	27	24	100
44	co2a0000422	22	25	27	74
45	co2a0000423	35	19	23	77
46	co2a0000424	43	30	29	102
47	co2a0000425	7	12	10	29
48	co2a0000426	32	17	20	69
49	co2a0000427	50	20	19	89
50	co2a0000428	39	19	20	78
51	co2a0000429	29	15	15	59
52	co2a0000430	55	26	27	108
53	co2a0000432	52	28	23	103
54	co2a0000433	47	17	15	79
55	co2a0000434	38	18	18	74
56	co2a0000435	39	20	19	78
57	co2a0000436	52	28	29	109
58	co2a0000437	53	29	29	111
59	co2a0000438	54	24	23	101
60	co2a0000439	48	28	26	102
61	co2a0000440	46	25	21	92

Table 4.9: Continuous from Table (4.8)

idx	subjects	# Trials with	# Trials with	# Trials with	Total # of Trials
		Experiment I	Experiment II	Experiment III	
62	co2a0000443	44	24	15	83
63	co2a0000444	42	24	14	80
64	co2a0000445	53	25	22	100
65	co2a0000447	43	28	21	92
66	co2c0000337	40	23	15	78
67	co2c0000338	54	27	23	104
68	co2c0000339	39	19	23	81
69	co2c0000340	54	12	17	83
70	co2c0000341	51	23	25	99
71	co2c0000342	56	29	30	115
72	co2c0000344	50	20	23	93
73	co2c0000345	55	28	29	112
74	co2c0000346	53	30	28	111
75	co2c0000347	46	21	22	89
76	co2c0000348	36	23	13	72
77	co2c0000351	50	18	24	92
78	co2c0000352	17	27	17	61
79	co2c0000354	47	20	25	92
80	co2c0000355	37	15	15	67
81	co2c0000356	49	27	22	98
82	co2c0000357	41	24	23	88
83	co2c0000359	33	18	18	69
84	co2c0000362	45	18	19	82
85	co2c0000363	55	27	28	110
86	co2c0000364	47	24	27	98
87	co2c0000367	51	26	25	102
88	co2c0000370	56	27	24	107
89	co2c0000371	33	16	10	59
90	co2c0000373	37	12	11	60
91	co2c0000374	51	28	27	106
92	co2c0000378	43	15	20	78
93	co2c0000379	50	26	24	100
94	co2c0000381	52	26	24	102
95	co2c0000382	54	28	27	109
96	co2c0000383	49	30	26	105
97	co2c0000384	42	22	24	88
98	co2c0000387	53	23	22	98
99	co2c0000388	24	24	18	66
100	co2c0000389	51	23	25	99
101	co2c0000390	18	20	21	59
102	co2c0000391	12	11	18	41
103	co2c0000392	51	24	24	99
104	co2c0000393	50	29	30	109
105	co2c0000394	48	18	19	85
106	co2c0000395	51	25	28	104
107	co2c0000396	28	18	11	57
108	co2c0000397	44	18	16	78
109	co2c1000367	53	24	26	103
110	co3a0000448	32	20	14	66
111	co3a0000450	60	29	30	119
112	co3a0000451	60	28	29	117
113	co3a0000453	48	18	20	86
114	co3a0000454	40	13	15	68
115	co3a0000455	27	16	15	58
116	co3a0000456	40	18	22	80
117	co3a0000457	33	15	23	71
118	co3a0000458	47	27	27	101
119	co3a0000459	53	20	25	98
120	co3a0000460	59	30	27	116
121	co3a0000461	13	29	29	71
122	co3c0000402	54	30	27	111

BIBLIOGRAPHY

BIBLIOGRAPHY

Abraham, A., M. P. Milham, A. Di Martino, R. C. Craddock, D. Samaras, B. Thirion, and G. Varoquaux (2017), Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example, *NeuroImage*, *147*, 736–745.

Adeli-Mosabbeb, E., K.-H. Thung, L. An, F. Shi, and D. Shen (2015), Robust feature-sample linear discriminant analysis for brain disorders diagnosis, in *Advances in Neural Information Processing Systems*, pp. 658–666.

Aronszajn, N. (1944), La théorie générale des noyaux réproduisants et ses applications, *Proceedings of the Cambridge Philosophical Society*, *39*, 133–153.

Banerjee, A., D. B. Dunson, and S. T. Tokdar (2012), Efficient gaussian process regression for large datasets, *Biometrika*, p. ass068.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008), Gaussian predictive process models for large spatial data sets, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(4), 825–848.

Belmokhtar, N., and N. Benamrane (2012), Classification of alzheimer's disease from 3d structural mri data, *Age*, *78*(8), 69–96.

Bennett, K. P., and C. Campbell (2000), Support vector machines: hype or hallelujah?, *ACM SIGKDD Explorations Newsletter*, *2*(2), 1–13.

Bhatkoti, P., and M. Paul (2016), Early diagnosis of alzheimer's disease: A multi-class deep learning framework with modified k-sparse autoencoder classification, in *Image and Vision Computing New Zealand (IVCNZ), 2016 International Conference on*, pp. 1–5, IEEE.

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992), A training algorithm for optimal margin classifiers, in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, ACM.

Bowman, F. D. (2005), Spatio-temporal modeling of localized brian activity, *Biostatistics*, *6*, 558–575.

Bowman, F. D., B. Caffo, S. S. Bassett, and C. Kilts (2008), A bayesian hierarchical framework for spatial modeling of fmri data, *NeuroImage*, *39*(1), 146–156.

Breiman, L. (1996), Bagging predictors, *Machine learning*, *24*(2), 123–140.

Breiman, L. (1999), Prediction games and arcing algorithms, *Neural computation*, *11*(7), 1493–1517.

Brosch, T., R. Tam, A. D. N. Initiative, et al. (2013), Manifold learning of brain mris by deep learning, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 633–640, Springer.

Chang, C.-C., and C.-J. Lin (2011), Libsvm: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.

Chang, Y.-W., C.-J. Hsieh, K.-W. Chang, M. Ringgaard, and C.-J. Lin (2010), Training and testing low-degree polynomial data mappings via linear svm, *Journal of Machine Learning Research*, *11*(Apr), 1471–1490.

Chen, S., and F. DuBois Bowman (2011), A novel support vector classifier for longitudinal high-dimensional data and its application to neuroimaging data, *Statistical Analysis and Data Mining*, *4*(6), 604–611.

Collobert, R., and S. Bengio (2001), Svmtorch: Support vector machines for large-scale regression problems, *The Journal of Machine Learning Research*, *1*, 143–160.

Crainiceanu, C. M., A.-M. Staicu, and C.-Z. Di (2009), Generalized multilevel functional regression, *Journal of the American Statistical Association*, *104*(488), 1550–1561.

Crammer, K., and Y. Singer (2002), On the learnability and design of output codes for multiclass problems, *Machine learning*, *47*(2-3), 201–233.

Dass, C., and J. S. Brodbelt (2001), Principles and practice of biological mass spectrometry, *Applied Spectroscopy*, *55*, 296.

Demiriz, A., K. P. Bennett, and J. Shawe-Taylor (2002), Linear programming boosting via column generation, *Machine Learning*, *46*(1-3), 225–254.

Developmental, D. M. N. S. Y., . P. Investigators, et al. (2014), Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010., *Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002)*, *63*(2), 1.

Di, C.-Z., C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi (2009), Multilevel functional principal component analysis, *The annals of applied statistics*, *3*(1), 458.

Di Martino, A., et al. (2014), The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, *Molecular psychiatry*, *19*(6), 659–667.

Duin, R. P., and D. M. Tax (2000), Experiments with classifier combining rules, in *Multiple classifier systems*, pp. 16–29, Springer.

Efron, B. (2005), Local false discovery rates.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004), Least angle regression, *The Annals of statistics*, *32*(2), 407–499.

Evans, A. C., D. L. Collins, S. Mills, E. Brown, R. Kelly, and T. M. Peters (1993), 3d statistical neuroanatomical models from 305 mri volumes, in *Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record.*, pp. 1813–1817, IEEE.

Evans, K. C., D. D. Dougherty, M. H. Pollack, and S. L. Rauch (2006), Using neuroimaging to predict treatment response in mood and anxiety disorders, *Annals of Clinical Psychiatry*, *18*(1), 33–42.

Fan, J., and Y. Fan (2008), High dimensional classification using features annealed independence rules, *Annals of statistics*, *36*(6), 2605.

Fein, D., et al. (2013), Optimal outcome in individuals with a history of autism, *Journal of child psychology and psychiatry*, *54*(2), 195–205.

Filippone, M., A. F. Marquand, C. R. Blain, S. C. Williams, J. Mourão-Miranda, and M. Girolami (2012), Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities, *The annals of applied statistics*, *6*(4), 1883.

Fitzmaurice, G. M. (1995), A caveat concerning independence estimating equations with multivariate binary data, *Biometrics*, pp. 309–317.

French, B. M., M. R. Dawson, and A. R. Dobbs (1997), Classification and staging of dementia of the alzheimer type: a comparison between neural networks and linear discriminant analysis, *Archives of neurology*, *54*(8), 1001–1009.

Freund, Y., and R. E. Schapire (1995), A desicion-theoretic generalization of on-line learning and an application to boosting, in *European conference on computational learning theory*, pp. 23–37, Springer.

Frie, T.-T., N. Cristianini, and C. Campbell (1998), The kernel-adatron algorithm: a fast and simple learning procedure for support vector machines, in *Machine Learning: Proceedings of the Fifteenth International Conference (ICML'98)*, pp. 188–196, Citeseer.

Friedman, J., T. Hastie, R. Tibshirani, et al. (2000), Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), *The annals of statistics*, *28*(2), 337–407.

Friston, K. J., J.-B. Poline, A. P. Holmes, C. D. Frith, and R. S. Frackowiak (1996), A multivariate analysis of pet activation studies, *Human brain mapping*, *4*(2), 140–151.

Friston, K. J., D. E. Glaser, R. N. Henson, S. Kiebel, C. Phillips, and J. Ashburner (2002), Classical and bayesian inference in neuroimaging: applications, *Neuroimage*, *16*(2), 484–512.

Gehler, P., and S. Nowozin (2009), On feature combination for multiclass object classification, in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 221–228, IEEE.

Gelfand, A. E., H.-J. Kim, C. Sirmans, and S. Banerjee (2003), Spatial modeling with spatially varying coefficient processes, *Journal of the American Statistical Association*, *98*(462), 387–396.

Greicius, M. D., B. Krasnow, A. L. Reiss, and V. Menon (2003), Functional connectivity in the resting brain: a network analysis of the default mode hypothesis, *Proceedings of the National Academy of Sciences*, *100*(1), 253–258.

Grove, A. J., and D. Schuurmans (1998), Boosting in the limit: Maximizing the margin of learned ensembles, in *AAAI/IAAI*, pp. 692–699.

Guo, R., M. Ahn, and H. Z. Hongtu Zhu (2015), Spatially weighted principal component analysis for imaging classification, *Journal of Computational and Graphical Statistics*, *24*(1), 274–296.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik (2002), Gene selection for cancer classification using support vector machines, *Machine learning*, *46*(1), 389–422.

Higgins, J. (1996), Brain imaging in psychiatry, chapter normal brain anatomy imaged by ct and mri, *Oxford: Blackwell Science. vector machines for temporal classification of block design fMRI data. NeuroImage*, *26*, 317–329.

Hinrichs, C., V. Singh, G. Xu, S. C. Johnson, A. D. N. Initiative, et al. (2011), Predictive markers for ad in a multi-modality framework: an analysis of mci progression in the adni population, *Neuroimage*, *55*(2), 574–589.

Honey, C., O. Sporns, L. Cammoun, X. Gigandet, J.-P. Thiran, R. Meuli, and P. Hagmann (2009), Predicting human resting-state functional connectivity from structural connectivity, *Proceedings of the National Academy of Sciences*, *106*(6), 2035–2040.

Hosseini-Asl, E., R. Keynton, and A. El-Baz (2016), Alzheimer's disease diagnostics by adaptation of 3d convolutional network, in *Image Processing (ICIP), 2016 IEEE International Conference on*, pp. 126–130, IEEE.

House, L. L., M. A. Clyde, R. L. Wolpert, et al. (2011), Bayesian nonparametric models for peak identification in maldi-tof mass spectroscopy, *The Annals of Applied Statistics*, *5*(2B), 1488–1511.

Hsu, C.-W., and C.-J. Lin (2002), A comparison of methods for multiclass support vector machines, *Neural Networks, IEEE Transactions on*, *13*(2), 415–425.

Hung, H., and C.-C. Wang (2012), Matrix variate logistic regression model with application to eeg data, *Biostatistics*, *14*(1), 189–202.

Joachims, T. (1998), Making large-scale svm learning practical, *Tech. rep.*, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund.

Kang, J., T. D. Johnson, T. E. Nichols, and T. D. Wager (2011), Meta analysis of functional neuroimaging data via bayesian spatial point processes, *Journal of the American Statistical Association*, *106*(493), 124–134.

Kang, J., W. Ye, L. Wang, A. Veiga-Lopez, V. Padmanabhan, and P. X. Song (2012), Local mixed-effects fitting for detecting reproductive hormone surge times, *Statistics in BioSciences*, *4*(2), 245–261.

Kang, J., B. J. Reich, and A.-M. Staicu (2016), Scalar-on-image regression via the soft-thresholded gaussian process, *arXiv preprint arXiv:1604.03192*.

Kecman, V., and I. Hadzic (2000), Support vectors selection by linear programming, in *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, vol. 5, pp. 193–198, IEEE.

Keerthi, S. S., S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy (2001), Improvements to platt's smo algorithm for svm classifier design, *Neural Computation*, *13*(3), 637–649.

Kerrouche, N., K. Herholz, R. Mielke, V. Holthoff, and J.-C. Baron (2006), 18fdg pet in vascular dementia: differentiation from alzheimer's disease using voxel-based multivariate analysis, *Journal of Cerebral Blood Flow & Metabolism*, *26*(9), 1213–1221.

Kittler, J., M. Hatef, R. P. Duin, and J. Matas (1998), On combining classifiers, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *20*(3), 226–239.

Kohannim, O., et al. (2010), Boosting power for clinical trials using classifiers based on multiple biomarkers, *Neurobiology of aging*, *31*(8), 1429–1442.

Kreßel, U. H.-G. (1999), Pairwise classification and support vector machines, in *Advances in kernel methods*, pp. 255–268, MIT press.

Lai, T. L., and D. Small (2007), Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(1), 79–99.

Li, B., M. K. Kim, and N. Altman (2010a), On dimension folding of matrix-or array-valued statistical objects, *The Annals of Statistics*, pp. 1094–1121.

Li, Y., H. Zhu, Y. Chen, H. An, J. Gilmore, W. Lin, and D. Shen (2009), Lstgee: Longitudinal analysis of neuroimaging data, in *SPIE Medical Imaging*, pp. 72,590F–72,590F, International Society for Optics and Photonics.

Li, Y., X. Lin, and P. Müller (2010b), Bayesian inference in semiparametric mixed models for longitudinal data, *Biometrics*, *66*(1), 70–78.

Lindquist, M. A. (2008), The statistical analysis of fmri data, *Statistical Science*, *23*(4), 439–464.

Liu, Y.-C., C.-C. K. Lin, J.-J. Tsai, and Y.-N. Sun (2013), Model-based spike detection of epileptic eeg data, *Sensors*, *13*(9), 12,536–12,547.

Magnin, B., L. Mesrob, S. Kinkingnéhun, M. Pélégrini-Issac, O. Colliot, M. Sarazin, B. Dubois, S. Lehéricy, and H. Benali (2009), Support vector machine-based classification of alzheimers disease from whole-brain anatomical mri, *Neuroradiology*, *51*(2), 73–83.

Marquand, A., M. Howard, M. Brammer, C. Chu, S. Coen, and J. Mourão-Miranda (2010), Quantitative prediction of subjective pain intensity from whole-brain fmri data using gaussian processes, *Neuroimage*, *49*(3), 2178–2189.

Morris, J. S., P. J. Brown, R. C. Herrick, K. A. Baggerly, and K. R. Coombes (2008), Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models, *Biometrics*, *64*(2), 479–489.

Ogawa, S., T.-M. Lee, A. R. Kay, and D. W. Tank (1990), Brain magnetic resonance imaging with contrast dependent on blood oxygenation, *Proceedings of the National Academy of Sciences*, *87*(24), 9868–9872.

Pal, M. (2008), Multiclass approaches for support vector machine based land cover classification, *arXiv preprint arXiv:0802.2411*.

Platt, J., et al. (1999), Fast training of support vector machines using sequential minimal optimization, *Advances in kernel methodssupport vector learning*, *3*.

Plis, S. M., et al. (2014), Deep learning for neuroimaging: a validation study, *Frontiers in neuroscience*, *8*.

Pollock, D. (2000), Trend estimation and de-trending via rational square-wave filters, *Journal of Econometrics*, *99*(2), 317–334.

Quinonero-Candela, J., and C. E. Rasmussen (2005), Analysis of some methods for reduced rank gaussian process regression, *Lecture Notes in Computer Science*, *3355*, 98–127.

Raghuraman, M., et al. (2001), Replication dynamics of the yeast genome, *science*, *294*(5540), 115–121.

Rangaswamy, M., and B. Porjesz (2014), Understanding alcohol use disorders with neuroelectrophysiology, *Handbook of clinical neurology*, *125*, 383.

Reiss, P. T., and R. T. Ogden (2010), Functional generalized linear models with images as predictors, *Biometrics*, *66*(1), 61–69.

Rogers, S. J., and L. A. Vismara (2008), Evidence-based comprehensive treatments for early autism, *Journal of Clinical Child & Adolescent Psychology*, *37*(1), 8–38.

Saffari, A., M. Godec, T. Pock, C. Leistner, and H. Bischof (2010), Online multi-class lpboost, in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3570–3577, IEEE.

Savio, A., M. García-Sebastián, C. Hernández, M. Graña, and J. Villanúa (2009), Classification results of artificial neural networks for alzheimers disease detection, in *Intelligent Data Engineering and Automated Learning-IDEAL 2009*, pp. 641–648, Springer.

Shi, Y., H.-I. Suk, Y. Gao, and D. Shen (2014), Joint coupled-feature representation and coupled boosting for ad diagnosis, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2721–2728.

Skup, M., H. Zhu, and H. Zhang (2012), Multiscale adaptive marginal analysis of longitudinal neuroimaging data with time-varying covariates, *Biometrics*, *68*(4), 1083–1092.

Song, P. X.-K., X. Gao, R. Liu, and W. Le (2006), Nonparametric inference for local extrema with application to oligonucleotide microarray data in yeast genome, *Biometrics*, *62*(2), 545–554.

Suk, H.-I., S.-W. Lee, D. Shen, A. D. N. Initiative, et al. (2015), Latent feature representation with stacked auto-encoder for ad/mci diagnosis, *Brain Structure and Function*, *220*(2), 841–859.

Sullivan Pepe, M., and G. L. Anderson (1994), A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data, *Communications in Statistics-Simulation and Computation*, *23*(4), 939–951.

Tzourio-Mazoyer, N., B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot (2002), Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain, *Neuroimage*, *15*(1), 273–289.

Vapnik, V. (1963), Pattern recognition using generalized portrait method, *Automation and remote control*, *24*, 774–780.

Vapnik, V. (1995), *The nature of statistical learning theory*, Springer Verlag, New York.

Vapnik, V. (1998), Statistical learning theory wiley-interscience, *New York*.

Veiga-Lopez, A., W. Ye, D. Phillips, C. Herkimer, P. Knight, and V. Padmanabhan (2008), Developmental programming: deficits in reproductive hormone dynamics and ovulatory outcomes in prenatal, testosterone-treated sheep, *Biology of reproduction*, *78*(4), 636–647.

Wang, X., B. Nan, J. Zhu, and R. Koeppe (2014), Regularized 3d functional regression for brain image data via haar wavelets, *The annals of applied statistics*, *8*(2), 1045.

Weston, J., and C. Watkins (1998), Multi-class support vector machines, *Tech. rep.*, Citeseer.

Weston, J., C. Watkins, et al. (1999), Support vector machines for multi-class pattern recognition., in *ESANN*, vol. 99, pp. 219–224.

Woods, R. P., S. T. Grafton, C. J. Holmes, S. R. Cherry, and J. C. Mazziotta (1998), Automated image registration: I. general methods and intrasubject, intramodality validation, *Journal of computer assisted tomography*, *22*(1), 139–152.

Worsley, K. J., and K. J. Friston (1995), Analysis of fmri time-series revisitedagain, *Neuroimage*, *2*(3), 173–181.

Wu, Q., and D.-X. Zhou (2005), Svm soft margin classifiers: linear programming versus quadratic programming, *Neural computation*, *17*(5), 1160–1187.

Xu, L., A. Krzyżak, and C. Y. Suen (1992), Methods of combining multiple classifiers and their applications to handwriting recognition, *Systems, man and cybernetics, IEEE transactions on*, *22*(3), 418–435.

Xu, S., M. Styner, J. Gilmore, J. Piven, and G. Gerig (2008), Multivariate nonlinear mixed model to analyze longitudinal image data: Mri study of early brain development, in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pp. 1–8, IEEE.

Zeger, S. L., and K.-Y. Liang (1986), Longitudinal data analysis for discrete and continuous outcomes, *Biometrics*, pp. 121–130.

Zhang, D., Y. Wang, L. Zhou, H. Yuan, D. Shen, A. D. N. Initiative, et al. (2011), Multimodal classification of alzheimer's disease and mild cognitive impairment, *Neuroimage*, *55*(3), 856–867.

Zhang, X. L., H. Begleiter, B. Porjesz, W. Wang, and A. Litke (1995), Event related potentials during object recognition tasks, *Brain Research Bulletin*, *38*(6), 531–538.

Zhao, Y., R. T. Ogden, and P. T. Reiss (2012), Wavelet-based lasso in functional linear regression, *Journal of Computational and Graphical Statistics*, *21*(3), 600–617.

Zhou, B., J. F. Xiao, L. Tuli, and H. W. Ressom (2012), Lc-ms-based metabolomics, *Molecular BioSystems*, *8*(2), 470–481.

Zhou, H., and L. Li (2014), Regularized matrix regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(2), 463–483.

Zhu, X., H.-I. Suk, and D. Shen (2014), A novel matrix-similarity based loss function for joint regression and classification in ad diagnosis, *NeuroImage*, *100*, 91–105.

Zou, H., and T. Hastie (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Zou, Q.-H., C.-Z. Zhu, Y. Yang, X.-N. Zuo, X.-Y. Long, Q.-J. Cao, Y.-F. Wang, and Y.-F. Zang (2008), An improved approach to detection of amplitude of low-frequency fluctuation (alff) for resting-state fmri: fractional alff, *Journal of neuroscience methods*, *172*(1), 137–141.